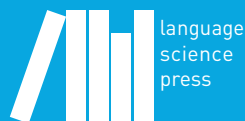


# Handbook of Lexical Functional Grammar

Edited by

Mary Dalrymple

Empirically Oriented Theoretical  
Morphology and Syntax



## Empirically Oriented Theoretical Morphology and Syntax

Chief Editor: Stefan Müller

Consulting Editors: Berthold Crysmann, Laura Kallmeyer

In this series:

1. Lichte, Timm. Syntax und Valenz: Zur Modellierung kohärenter und elliptischer Strukturen mit Baumadjunktionsgrammatiken.
2. Bîlbîie, Gabriela. Grammaire des constructions elliptiques: Une étude comparative des phrases sans verbe en roumain et en français.
3. Bowerman, Claire, Laurence Horn & Raffaella Zanuttini (eds.). On looking into words (and beyond): Structures, Relations, Analyses.
4. Bonami, Olivier, Gilles Boyé, Georgette Dal, Hélène Giraudo & Fiammetta Namer. The lexeme in descriptive and theoretical morphology.
5. Guzmán Naranjo, Matías. Analogical classification in formal grammar.
6. Flick, Johanna. Die Entwicklung des Definitartikels im Althochdeutschen: Eine kognitiv-linguistische Korpusuntersuchung.
7. Zinova, Yulia. Russian verbal prefixation: A frame semantic analysis.

# Handbook of Lexical Functional Grammar

Edited by

Mary Dalrymple


Mary Dalrymple (ed.). 2021. *Handbook of Lexical Functional Grammar* (Empirically Oriented Theoretical Morphology and Syntax). Berlin: Language Science Press.

This title can be downloaded at:

<http://langsci-press.org/catalog/book/312>

© 2021, the authors

Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):

<http://creativecommons.org/licenses/by/4.0/> 

ISBN: no digital ISBN

no print ISBNs!

ISSN: 2366-3529

no DOI

Source code available from [www.github.com/langsci/312](http://www.github.com/langsci/312)

Collaborative reading: [paperhive.org/documents/remote?type=langsci&id=312](http://paperhive.org/documents/remote?type=langsci&id=312)

Cover and concept of design: Ulrike Harbort

Fonts: Libertinus, Arimo, DejaVu Sans Mono, Source Han Serif ZH, Source Han Serif JA

Typesetting software: Xe<sub>La</sub>T<sub>E</sub>X

Language Science Press

xHain

Grünberger Str. 16

10243 Berlin, Germany

[langsci-press.org](http://langsci-press.org)

Storage and cataloguing done by FU Berlin

Freie Universität



Berlin

# Contents

## I Overview and introduction

- 1 Introduction to LFG  
Oleg Belyaev 3
- 2 Core concepts of LFG  
Oleg Belyaev 21

## II Grammatical modules and interfaces

- 3 Prosody and its interfaces  
Tina Bögel 95

## III Formal and computational issues and applications

- 4 Computational implementations and applications  
Martin Forst & Tracy Holloway King 141
- 5 Treebank-driven Parsing, Translation and Grammar Induction using  
LFG  
Aoife Cahill & Andy Way 181

## IV Language families and regions

- 6 LFG and Finno-Ugric languages  
Tibor Laczkó 227
- 7 LFG and Romance languages  
Alex Alsina 291

*Contents*

<b>8</b>	<b>LFG and Semitic languages</b>	
	Louisa Sadler	<b>355</b>
<b>V</b>	<b>Comparing LFG with other linguistic theories</b>	
<b>9</b>	<b>LFG and HPSG</b>	
	Adam Przepiórkowski	<b>409</b>
	<b>Index</b>	<b>467</b>

## **Part I**

# **Overview and introduction**





# Chapter 1

## Introduction to LFG

Oleg Belyaev

Lomonosov Moscow State University and Institute of Linguistics of the Russian Academy of Sciences

This chapter provides a general summary of the architecture of LFG. It is mainly focused on describing the two main syntactic levels, c- and f-structure, and the projection architecture used in LFG in general. It also describes the notation for defining the range of possible c-structures and their corresponding f-structures. Core syntactic mechanisms such as structure sharing and X-bar theory are also briefly covered.

### 1 Introduction

This chapter aims to summarize the main syntactic levels of LFG, constituent structure (c-structure) and functional structure (f-structure), while providing a general overview of the foundational features of this framework. In Section 2, I briefly describe the basic architecture of LFG and the overall role played by each of the syntactic levels. In Section 3, I describe the c-structure model used in standard LFG, its understanding of constituency, and the role of X' theory. In Section 4, the notion of f-structure is introduced, together with notational conventions and a system of mapping c-structure to f-structure. In Section 5, I show how the basic system of c- and f-structure can be extended to include other levels of projection that comprise the architecture of LFG.

### 2 The basic architecture of LFG

At the core of LFG architecture as it was originally proposed in Kaplan & Bresnan (1982) is the split of syntax into two levels: constituent structure, or **c-structure**,

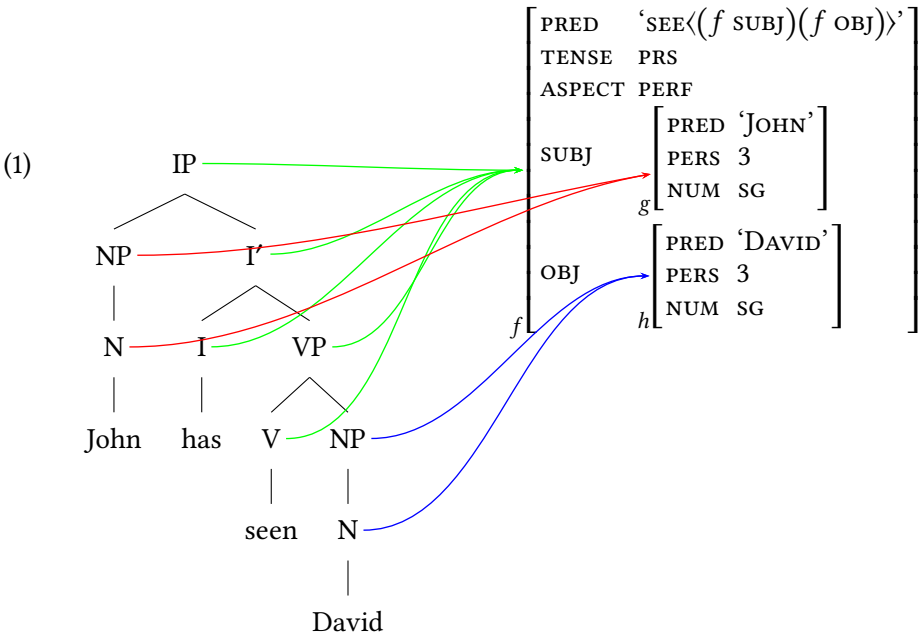


Oleg Belyaev. 2021. Introduction to LFG. in Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 3–20. Berlin: Language Science Press. DOI:

?? 

Oleg Belyaev

and functional structure, or **f-structure**. The correspondence function  $\phi(x)$  maps every c-structure node to an f-structure. As an example, consider the LFG analysis of the sentence *John has seen David* in (1), where the mapping function is represented by the arrows.



As seen in (1), the two parallel structures are substantially different: c-structure is a phrase structure tree that represents word order and hierarchical embedding, while f-structure is a feature-value structure that represents predicate-argument relations and the grammatical features of all the major parts of the sentence. Features appear as atomic values of f-structure attributes, while arguments and adjuncts appear as f-structures embedded as values of attributes such as SUBJ and OBJ in (1); which arguments can and, indeed, have to appear in the f-structure is specified in the value of the PRED attribute. While the mapping between the two structures follows certain constraints imposed both by the formal metalanguage and theoretical considerations (on which see Belyaev 2021 [this volume]), it is, in principle, language-specific: an LFG grammar consists of a set of rules and lexical entries that define the possible c-structures and their corresponding f-structures for a particular language.

This flexibility in the c- to f-structure correspondence ensures that each corresponds to a particular set of grammatical generalizations. Overall, f-structure is

the main syntactic level that represents the predicates, their valencies and grammatical relations, as well as grammatical features such as number, case, aspect and gender. The majority of syntactic phenomena that have to do with feature assignment and feature checking are described using f-structure constraints; these include:

- feature government (case assignment, mood, constraints on the use of non-finite forms, etc.);
- agreement;
- anaphoric constraints;
- wh-movement, topicalization and other long-distance dependencies.

All generalizations that have to do with argument relations and grammatical features have to be stated in terms of f-structure. For instance, a constraint that requires the verb to agree with Spec,IP or to assign accusative case to Comp,VP would be complex and somewhat unnatural to formulate (although not impossible). It is much more simple and natural in LFG for such rules to refer to grammatical functions such as SUBJ and OBJ instead. This implies that the role of constituent structure is more restricted than in other frameworks; for the most part, c-structure constraints only capture generalizations related to word order and various embedding possibilities.

The correspondence architecture is not limited to syntax. Other projections that map c-structure nodes or f-structures to other structures (such as information structure, semantic structure, or prosody) have been proposed in the literature: see Section 5 for details.

## 3 C-structure

### 3.1 The notion of c-structure

C-structure (constituent structure) in LFG is a phrase structure tree. Possible trees are defined by a set of context-free statements (“phrase structure rules”) of the type  $A \rightarrow \alpha$ , where  $A$  is a nonterminal symbol (representing some syntactic category), while  $\alpha$  is a string of nonterminals or a single terminal. A simple set of rules that licenses the English sentence in (1) is given in (2).

- (2)    a.  $IP \rightarrow NP \ I'$     b.  $I' \rightarrow I \ VP$     c.  $VP \rightarrow V \ NP$     d.  $NP \rightarrow N$   
       e.  $N \rightarrow \text{John}$     f.  $N \rightarrow \text{David}$     g.  $I \rightarrow \text{has}$     h.  $V \rightarrow \text{seen}$

*Oleg Belyaev*

Such rules are well-established in modern linguistics since at least Chomsky (1957) and so hardly require further discussion. It should however be observed that, in LFG, these should not be understood as “rules” in the direct (procedural) sense, but rather a set of phrase structure principles that constrain hierarchical relations between mothers and daughters – crucially, not between levels further apart, such as granddaughters etc. Phrase structure grammars are one way of describing such principles that has proved most popular among LFG practitioners, but not the only way – possible alternatives are ID/LP rules (Falk 1983) and the specification language described in Potts (2002), which builds on the specification language in Blackburn & Gardent (1995).

The structures that are constrained in this way are not just strings,<sup>1</sup> but constituent structure trees whose nodes are individually mapped to f-structures, as shown in (1).

The syntax of phrase structure rules in LFG is somewhat more extensive than in many other frameworks, because the right-hand side  $\alpha$  is allowed to be a regular expression and include such features as optionality (represented by parentheses around the symbol), disjunction (with the disjuncts in curly brackets, separated by either a vertical line (|) or a logical disjunction sign ( $\vee$ ): e.g. { NP | DP }), Kleene star (zero or more instances,  $\text{NP}^*$ ), Kleene plus (one or more instances,  $\text{NP}^+$ ), and some other less frequently used expressions. Grammars where the right-hand side can include regular expressions are called extended context-free grammars or regular right part grammars and it is known (Woods 1970) that the set of languages they describe is the same as that of standard context-free grammar.

### 3.2 Main properties of c-structure

LFG is unique among all frameworks in the simplicity of its constituent structure representations. This is a deliberate design decision which is possible due to the parallel architecture approach of LFG. It has been widely accepted since Chomsky (1957) that context-free grammar is not by itself an adequate formalism for describing natural language; even if the majority of syntactic constructions can indeed be described by context-free grammar (Pullum & Gazdar 1982), the descriptions required would be cumbersome, artificial and theoretically unenlightening as a model of human linguistic competence. Therefore, most grammars

---

<sup>1</sup>In fact, in the original version of LFG architecture introduced in Kaplan (1989), c-structure is itself a projection from the string. In recent LFG work, this idea has been developed in more detail by distinguishing between the *s-string* (the string of syntactic units) and the *p-string* (the string of phonological units), see Dalrymple & Mycock (2011) and Bögel 2021 [this volume] for more information.

## 1 Introduction to LFG

which use constituent structure as the main level of syntactic representation introduce additional mechanisms such as transformations in order to increase their expressive power. But such additions are not required in LFG because all phenomena that require more powerful mechanisms are dealt with at f-structure and other levels. C-structure remains limited to modeling basic word order facts, hierarchical embedding, and recursion, the phenomena for which phrase structure always was and remains the most adequate formal representation.

The advantage of this simplicity is that constituent structure in LFG has a clear empirical basis and can be determined for individual languages based on classic tests not obscured by additional considerations. For example, since there is no syntactic displacement, constituents in LFG are continuous by definition – apparently “discontinuous” material may eventually converge in one f-structure, but will still be split into separate constituents at c-structure.

By contrast, some constituency diagnostics which are valid in other frameworks are not valid in LFG. For example, since c-command is a phrase structure-based relation in mainstream transformational grammar, the existence of binding asymmetries between subjects and objects implies a configurational structure where the subject c-commands the object or vice versa. Thus Speas (1990: 137) argues that, within standard GB assumptions, flat structure predicts the existence of subject reflexives bound by their objects; since few such languages, if any, are actually found, existence of a hierarchical structure with a VP and a subject c-commanding the direct object is part of Universal Grammar.

In LFG, such a conclusion is a *non sequitur* because constraints on anaphoric relations, and other related phenomena, are formulated chiefly in terms of f-structure; sometimes in terms of information structure, semantics, or even linear precedence; but almost never in terms of c-structure configuration. Reference to c-command is possible in principle,<sup>2</sup> but it is largely useless as a source of valid generalizations due to the core assumptions of LFG: the cross-linguistic variability of c-structure, the universality<sup>3</sup> of grammatical functions at f-structure, and variation in the syntax-semantics interface.

Constituent structure representations in LFG are therefore rather “shallow” in that their makeup is determined by a limited set of empirical diagnostics mostly based on word order possibilities. These facts vary widely across languages, and

---

<sup>2</sup>As, for example, in the definition of extended heads in Bresnan et al. (2016: 136). Note that this is a concept that is used to describe regularities in the c- to f-structure mapping, not a constraint on f-structure relations themselves.

<sup>3</sup>“Universality” here refers to universal availability, as in a grammatical toolbox (cf. Jackendoff 2002), not in the sense of mapping the same semantic roles to the same grammatical functions in all languages, or even in a single language. See **chapters/GFs** for more detail.

*Oleg Belyaev*

so do c-structure rules and the resulting structures. While f-structures have a degree of universality (in the sense of sharing a single inventory of grammatical functions and broad similarity in the way analogous phenomena such as anaphora, coordination, agreement etc. are represented), c-structures are language-specific.

Still, even in c-structure there are certain basic theoretical constraints which are deemed to hold universally across languages. In mainstream LFG, these are ENDOCENTRICITY and LEXICAL INTEGRITY. The former is usually captured by a version of X-Bar Theory, which is generally the same as in GB (see Chomsky 1970; Jackendoff 1977) but less restrictive: no universal clause or NP structure, no universal mapping from  $X'$ -theoretic positions (specifier, complement) to grammatical functions are assumed; non-binary branching is allowed; various exceptions from endocentricity, most prominently the exocentric S node used in non-configurational languages are permitted. For more information on the version of X-Bar Theory used in LFG, see Belyaev 2021 [this volume].

Lexical integrity is another principle that has been assumed in LFG since its inception. At its core, this principle states that words are constructed from different elements and according to different rules than syntactic phrases, and that the internal structure of words is invisible to rules of syntax (Bresnan & Mchombo 1995: 181). In formal terms, this is usually interpreted such that the leaves of c-structure trees must be morphologically complete words (Bresnan et al. 2016: 92). For more detail on lexical integrity as it is used in LFG, the challenges it faces and proposed modifications, see Belyaev 2021 [this volume].

## 4 F-structure

### 4.1 Defining equations

As mentioned above, at the most basic level f-structures in LFG are a type of attribute-value structure.<sup>4</sup> However, unlike most other frameworks which deal with this data type, the LFG formalism does not refer to f-structures as objects that can be manipulated and to which various operations can be applied. In contrast, an f-structure is thought of as a *function* that maps attributes (attribute names) to their values.<sup>5</sup>

---

<sup>4</sup>Carpenter (1992) is the standard reference on the mathematical properties of such feature structures. However, the structures described by Carpenter are *typed*, which is a crucial difference from LFG f-structures, which are *untyped* and defined using a functional notation.

<sup>5</sup>The term *f(unctional)-structure* can thus be understood in two ways: as a structure representing the “function” of words and phrases (as opposed to c-structure which represents “form”) and,

## 1 Introduction to LFG

From this perspective, describing an f-structure consists in defining the value  $y$  for each argument  $x$  in the function's domain (i.e. the set of attribute names). In LFG, attribute-value pairs are usually described using the notation of function application probably inspired by the Lisp programming language, i.e. the more conventional  $f(x) = y$  is expressed as  $(f\ x) = y$ . Thus, for the f-structure  $f$  in (1), the value of the attribute TENSE is defined by the equation  $(f\ \text{TENSE}) = \text{PRS}$ . By way of example, the full (minimal) set of equations that describes the f-structure of (1) is provided in (3).

- (3)  $(f\ \text{PRED}) = \text{'SEE'}$   
 $(f\ \text{TENSE}) = \text{PRS}$   
 $(f\ \text{ASPECT}) = \text{PERF}$   
 $(f\ \text{SUBJ}) = g$   
 $(f\ \text{OBJ}) = h$   
  
 $(g\ \text{PRED}) = \text{'JOHN'}$   
 $(g\ \text{PERS}) = 3$   
 $(g\ \text{NUM}) = \text{SG}$   
  
 $(g\ \text{PRED}) = \text{'DAVID'}$   
 $(g\ \text{PERS}) = 3$   
 $(g\ \text{NUM}) = \text{SG}$

Sets of equations as in (3) are called F-DESCRIPTIONS. A valid f-structure of a sentence is an f-structure that *minimally* satisfies this sentence's f-description. Thus, the f-structure displayed in (1) is the minimal f-structure that satisfies (3); were one to add the attribute-value pair [MOOD INDICATIVE], (3) would still be satisfied, but the structure would no longer be minimal.

Since an f-structure functional application produces attribute values, and, as seen in (1) and (3), these values can also be f-structures, it is possible to use nested function applications. Thus, since  $(f\ \text{SUBJ}) = g$ ,  $((f\ \text{SUBJ})\ \text{PERS})$  is equivalent to  $(g\ \text{PERS})$  and has the value 3. By convention, function application is left associative, thus the parentheses can be omitted and the equation written as  $(f\ \text{SUBJ}\ \text{PERS}) = 3$ . Early on, LFG has also adopted an extension of function application called functional uncertainty (Kaplan & Zaenen 1989), which allows replacing the right-hand side of the function application (the “path” of attribute names) by a regular expression; thus,  $(f\ \text{COMP}^*\ \text{SUBJ})$  denotes the value of the

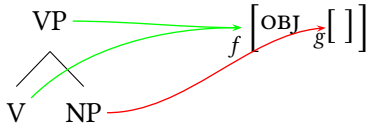
---

more formally, as a *function* proper. This set-theoretic understanding of f-structures is standard in the LFG literature, but f-structures can alternatively be modeled in terms of graph theory; an example of this approach is found in Kuhn (2003).





(6)



For convenience,  $\phi(*)$  and  $\phi(\hat{*})$  are usually replaced by the abbreviations  $\downarrow$  (pronounced “down”) and  $\uparrow$  (pronounced “up”), respectively. These metavariables are assumed to be the only way to refer to material up or down the tree in phrase structure rules; direct reference to “low-level” variables such as  $*$  is generally not used in LFG analyses. The conventional representation of the rule in (5) is given in (7).

$$(7) \quad \text{VP} \longrightarrow \begin{array}{cc} \text{V} & \text{NP} \\ \uparrow=\downarrow & (\uparrow \text{OBJ})=\downarrow \end{array}$$

In the standard model of c-structure, lexical entries are nothing more than rules defining a preterminal node dominating a terminal node. However, they use a slightly different notation, where the word form is followed by its category and annotation, illustrated in (8).

$$(8) \quad \textit{John} \quad \text{N} \quad \begin{array}{l} (\uparrow \text{PRED})=\text{'JOHN'} \\ (\uparrow \text{PERS})=3 \\ (\uparrow \text{NUM})=\text{SG} \end{array}$$

Since there is no further material down the tree, lexical entries typically only use the metavariable  $\uparrow$  to provide information associated with the preterminal node. In some cases,  $\downarrow$  is also used to draw subtle distinctions between information contributed by the word itself and the information contributed by the preterminal. For example, Zaenen & Kaplan (1995: 230) ingeniously map the verbal form to the *PRED* value, while other grammatical features are assumed to be contributed by the *V* node. In practice, this possibility is seldom used.

The projection function  $\phi$  maps c-structure nodes to f-structures, but one may also define an inverse correspondence  $\phi^{-1}$  to proceed in the opposite direction. This function provides the set of c-structure nodes that map to the f-structure given as its argument. Note that the inverse projection is not a function, as the c- to f-structure relation is one-to-many. Inverse projections are used in f-descriptions in order to use c-structure features in f-structure constraints. For example, to check that the subject’s f-structure maps to an NP, one may use the equation  $\text{NP} \in \text{CAT}((\uparrow \text{SUBJ})^{-1})$ . This is seldom needed, because, by design,

*Oleg Belyaev*

most constraints on f-structure attributes can be described solely in terms of f-structure. However, sometimes the inverse projection is indispensable, e.g. when formulating the notion of f-precedence (see Belyaev 2021 [this volume]) describing linear order conditions on anaphora (see **chapters/Anaphora**).

### 4.3 Well-formedness conditions

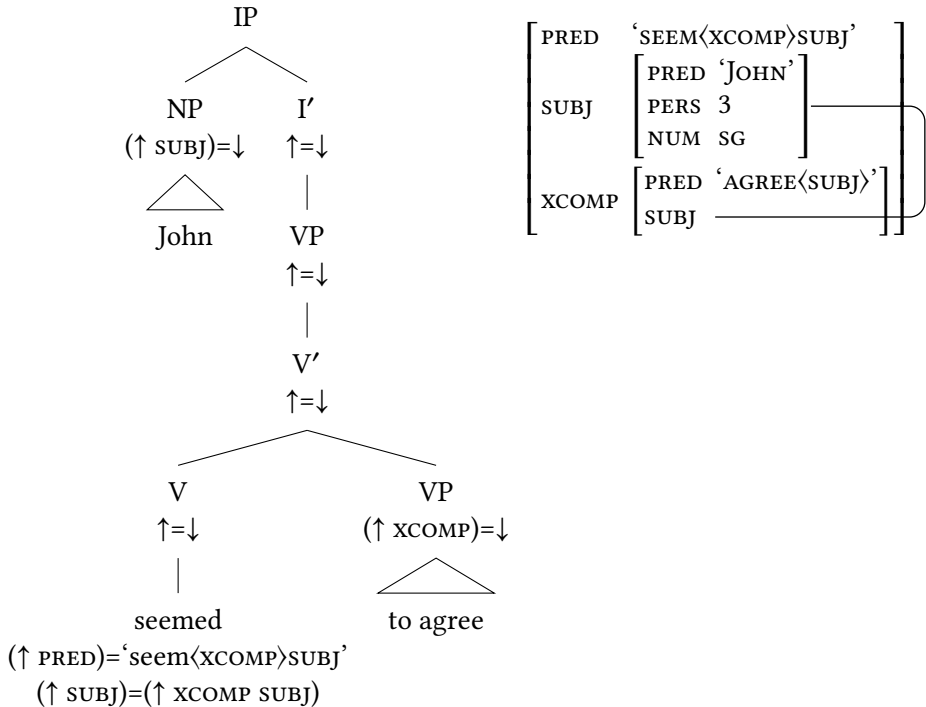
There are three conditions that any f-structure must satisfy in order to be treated as valid: Uniqueness (also known as Consistency), Completeness, and Coherence. Any f-structure that violates these conditions cannot be part of a valid analysis of any sentence, regardless of the grammar as a whole. Uniqueness requires that each attribute have exactly one value – this actually follows from the notion of f-structure as a function, since a function, by definition, is a many-to-one or one-to-one mapping. Completeness requires that each argument listed in the PRED value of an f-structure (which is the locus of valency information) is present in the f-structure; Coherence, complementarily, requires that no extra arguments *not* listed in the PRED value are introduced. For more detail on how these conditions actually operate, see Belyaev 2021 [this volume].

### 4.4 Structure sharing and “movement”

Unlike transformation-based grammatical approaches, LFG has no special formal mechanism such as movement or Internal Merge to handle dependencies between different structural positions. The closest equivalent to such a mechanism is STRUCTURE SHARING, which has already been mentioned above. The possibility of structure sharing follows from the general makeup of the formalism: If f-structures are functions and features are their arguments, it is expected that these structures are reentrant: a function can return the same value for different arguments. Since reentrancy is obviously required for the simplest cases such as reentrant atomic values, structure sharing is only a natural consequence of this property.

A classic example of the use of structure sharing to describe a movement-like process is the LFG analysis of raising. Raising verbs such as English *seem* are analyzed as having a non-thematic subject that is shared with the subject of the complement clause:

(9)



This correctly predicts that the raised subject appears as the argument of the matrix clause while being subcategorized for and assigned a semantic role in the complement clause. For more detail on control and raising, see **chapters/Control**.

It is important to note that while structure sharing is, in formal terms, the closest counterpart to movement in LFG, this does not mean that all phenomena that are treated via movement in transformational frameworks should involve structure sharing in LFG. This is because movement is normally the *only* mechanism for “non-canonical” or “displaced” positioning of material in transformational frameworks, while LFG draws a crucial distinction between c- and f-structure. Two sentences may differ in the c-structure while having the same f-structure – this is called **SCRAMBLING** and this is the most widespread mechanism of syntactic “displacement” in non-configurational languages or languages that allow mapping to the same grammatical function in different positions. For example, Arka (2003) proposes the following rule for S in Balinese:

$$(10) \quad S \quad \longrightarrow \quad \left\{ \begin{array}{c|c} \text{VP} & \text{NP} \\ \hline \uparrow=\downarrow & (\uparrow \text{GF})=\downarrow \end{array} \right\}^*$$

*Oleg Belyaev*

This allows any number of NPs to alternate with any number of VPs in any order; each NP may be freely assigned to any grammatical function. Therefore, sentences with the same predicate and the same set of NP arguments will have identical f-structures, with the only difference being found at c-structure. But no c-structure configuration will be considered as “basic” in any formal sense of the term.<sup>6</sup>

## 5 Additional levels of projection

C-structure and f-structure were originally thought of as the only levels of grammar in LFG: c-structure as a kind of “form” representation, and f-structure as a “functional” representation, in some sense reflecting semantics and having a degree of universality compared to c-structure. It quickly became clear, however, that these two levels are not enough to represent the full complexity of grammatical phenomena. First, semantics should be separate from f-structure to handle phenomena that are not represented in syntax, such as quantifier scope. Second, f-structure in its standard form is a collection of information of different types: purely morphological and morphosyntactic atomic features; grammatical functions; valency information (PRED features); and semantic information (if features such as ANIM are used to describe effects of animacy on grammatical marking). Third, f-structure simply cannot handle some phenomena, like prosody, which require a different kind of structure whose constituents are not equivalent to either c-structure constituents or f-structures.

A possible way to overcome these difficulties would be to extend the role of the existing c- and f-structure, which would mirror similar developments in transformational grammar, with its central role of constituent structure and the proliferation of functional projections. However, the architecture of LFG permits a more elegant solution. While the original system does only consist of c- and f-structure, there is nothing intrinsic about this binarity: the two are connected by a projection function  $\phi$  that maps nodes to f-structure. It is possible to define other functions that would connect c- or f-structures to various other structures; thus, where  $\phi(*)$  (abbreviated  $\downarrow$ ) stands for the f-structure of the annotated node,  $\mu(*)$  would be the morphosyntactic structure (m-structure) of this c-structure node,

---

<sup>6</sup>Of course, even in non-configurational languages, certain word orders are often viewed as less marked compared to others. This is probably due to differences in information structure, which in modern LFG literature is usually treated as a separate level that may interact with other levels such as c-structure, f-structure, and prosody (Dalrymple & Nikolaeva 2011). Crucially, an information structure difference between sentences does not automatically entail any difference at either c- or f-structure.

## 1 Introduction to LFG

and  $\sigma(\phi(*))$  (abbreviated  $\uparrow_\sigma$ ) would be the semantic structure (s-structure) that the f-structure that corresponds to this node maps to (if s-structure is viewed as projected from f-structure). The simultaneous description of two or more grammatical structures by the same rule or lexical entry is called CODESCRIPTION, which is the main principle governing the interaction of levels in LFG.

This modularity has been successfully used to model a number of grammatical levels, such that LFG, as it is currently practiced, is no longer centered around the interaction between c- and f-structure, although these still play a major role as the main syntactic representations. It is also crucial that LFG, by design, still retains a degree of “syntactocentricity” in that all additional projections are defined with reference to c-structure nodes. This is different from the notion of a truly parallel architecture advocated e.g. in Culicover & Jackendoff (2005), where each level of representation (specifically, in their model, syntax and semantics) is conceived of as a separate “combinatorially autonomous” system that is linked to other levels via a system of correspondence constraints. In LFG, only c-structure is combinatorial in this sense,<sup>7</sup> with possible trees defined directly through phrase structure rules; the content of other projections is not autonomously generated, but defined through phrase structure annotations that connect the elements of these projections to c-structure nodes. Thus, while c-structure is not as central as constituent structure in other frameworks, it acts as a “hub” that connects all the different levels of sentence structure together.<sup>8</sup>

There is currently no agreed-upon set of representational levels. Some, like s-structure or prosodic structure, are almost universally adopted and consistently interpreted in terms of projection. Others, like information structure (i-structure), are assumed by most authors, but specific interpretations vary: for example, i-structure is projected from c-structure in King (1997); Butt & King (1997), but from s-structure in the more recent proposal of Dalrymple & Nikolaeva (2011).

---

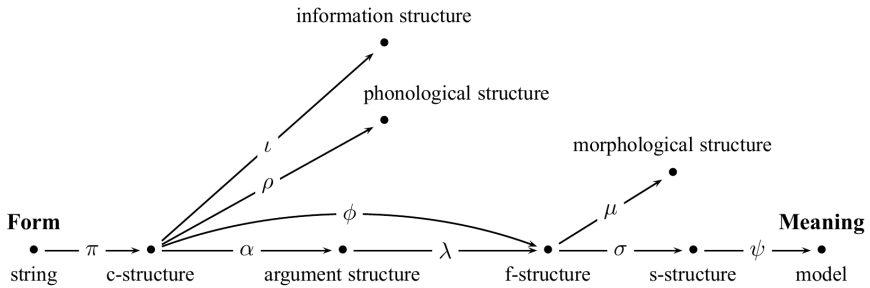
<sup>7</sup>C-structure rules are somewhat less central in approaches like Halvorsen (1983) and Andrews (2008), which use description by analysis, rather than the standard codescription approach, to describe the syntax-semantics interface: In these approaches, meaning is constructed on the basis of f-structure, without direct reference to c-structure. Even here, however, semantics is not a separate combinatorial system but is constructed on the basis of another structure which, in turn, is projected from c-structure; this still seems rather different from Culicover and Jackendoff’s vision of parallel architecture.

<sup>8</sup>This flavour of syntactocentricity is far less radical than in mainstream generative grammar and may in fact be unavoidable in a (broadly) lexicalist framework, inasmuch as words are viewed as the “building blocks” of sentences. In fact, I am not aware of a fully developed and formalized implementation of any truly parallel architecture. There is no way around the fact that phonetic form is the only part of language that is directly available for perception; thus the part of grammar that is tasked with combining such “surface” elements into complete utterances – i.e. syntax in the narrow sense – will always have a special role.

*Oleg Belyaev*

Finally, some levels are specific to particular approaches and are not universally adopted, e.g. morphosyntactic structure (m-structure), viewed as projected from c-structure (Butt et al. 2004; Butt, Fortmann, et al. 1996) or f-structure (Sadler & Nordlinger 2004); or argument structure (a-structure), which is used in some approaches to argument mapping (Butt et al. 1997) but is viewed as redundant in some more recent proposals such as (Asudeh & Giorgolo 2012; Asudeh et al. 2014; Findlay 2016). One version of how the correspondence architecture might look is provided in (11).

(11)



To date, the following additional levels and projections have been discussed and described in the LFG literature (references to some of the proposals are given in parentheses; most have separate chapters in the handbook, which describe proposed representations in detail):

- argument structure (a-structure) (Butt et al. 1997), see **chapters/Mapping**;
- semantic structure (s-structure) (Dalrymple 1999), see **chapters/Glue**;
- information structure (i-structure) (King 1997; Butt & King 1997; Dalrymple & Nikolaeva 2011), see **chapters/InformationStructure**;
- prosodic structure (p-structure) (Dalrymple & Mycock 2011; Bögel 2012), see Bögel 2021 [this volume];
- morphological / morphosyntactic structure (m-structure), see (Butt et al. 2004; Sadler & Nordlinger 2004), **chapters/Morphology**;
- grammatical marking structure (g-structure) (Falk 2006);
- l-structure, used in Lexical Sharing (Wescoat 2002, 2005).

## 6 Conclusion

In this chapter, I have described the main architectural notions of LFG – the c- and f-structures. LFG can be viewed as incorporating the best features of constituent-structure-based (at c-structure) and dependency-based (at f-structure) frameworks, while avoiding their main drawbacks. Frameworks that use phrase structure as the only syntactic representation require additional mechanisms such as transformations, multiple dominance or separate linearization to properly capture word order variation and feature constraints; LFG manages to keep c-structure relatively simple due to the fact that all feature interactions are captured at f-structure, without referring to constituent structure positions. At the same time, the fact that f-structure does not directly refer to individual words or phrase structure nodes allows adequately capturing word order variation while keeping predicate-argument representations fairly uniform across languages.

## Acknowledgements

This research has been supported by the Interdisciplinary Scientific and Educational School of Moscow University “Preservation of the World Cultural and Historical Heritage”.

## References

- Andrews, Avery D. 2008. The role of PRED in LFG+Glue. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 46–67. Stanford, CA: CSLI Publications.
- Arka, I Wayan. 2003. *Balinese morphosyntax: A lexical-functional approach*. Canberra: Pacific Linguistics.
- Asudeh, Ash & Gianluca Giorgolo. 2012. Flexible composition for optional and derived arguments. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 64–84. Stanford, CA: CSLI Publications.
- Asudeh, Ash, Gianluca Giorgolo & Ida Toivonen. 2014. Meaning and valency. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 68–88. Stanford, CA: CSLI Publications.
- Belyaev, Oleg. 2021. Core concepts of LFG. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 21–92. Berlin: Language Science Press. DOI: ??.

Oleg Belyaev

- Blackburn, Patrick & Claire Gardent. 1995. A specification language for Lexical-Functional Grammars. In *Proceedings of the 7th conference of the European chapter of the ACL (EACL 1995)*, 39–44. European Association for Computational Linguistics. DOI: 10.3115/976973.976980.
- Bögel, Tina. 2012. The p-diagram – a syllable-based approach to p-structure. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 99–117. Stanford, CA: CSLI Publications.
- Bögel, Tina. 2021. Prosody and its interfaces. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 95–137. Berlin: Language Science Press. DOI: ??.
- 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 & Sam A. Mchombo. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language & Linguistic Theory* 13(2). 181–254. DOI: 10.1007/bf00992782.
- Butt, Miriam, Mary Dalrymple & Anette Frank. 1997. An architecture for linking theory in LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '97 conference*, 1–16. Stanford, CA: CSLI Publications.
- Butt, Miriam, Christian Fortmann & Christian Rohrer. 1996. Syntactic analyses for parallel grammars. In *COLING '96: Proceedings of the 16th Conference on Computational Linguistics*, 182–187. Copenhagen: International Committee on Computational Linguistics. DOI: 10.3115/992628.992662.
- Butt, Miriam & Tracy Holloway King. 1997. Null elements in discourse structure. Unpublished manuscript intended for publication in the Proceedings of the NULLS Seminar. <http://ling.uni-konstanz.de/pages/home/butt/main/papers/nulls97.pdf>.
- Butt, Miriam, María-Eugenia Niño & Frédérique Segond. 2004. Multilingual processing of auxiliaries within LFG. In Louisa Sadler & Andrew Spencer (eds.), *Projecting morphology*. Stanford, CA: CSLI Publications. Previously published as Butt, Niño & Segond (1996).
- Butt, Miriam, María-Eugenia Niño & Frederique Segond. 1996. Multilingual processing of auxiliaries in LFG. In D. Gibbon (ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS conference*, 111–122. Berlin: Mouton de Gruyter.
- Carpenter, Robert L. 1992. *The logic of typed feature structures*. Cambridge. UK: Cambridge University Press. DOI: 10.1017/cbo9780511530098.
- Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton. DOI: 10.1515/9783110218329.



## 1 Introduction to LFG

- Chomsky, Noam. 1970. Remarks on nominalization. In Roderick A. Jacobs & Peter S. Rosenbaum (eds.), *Readings in English transformational grammar*, 184–221. Waltham, MA: Ginn.
- Culicover, Peter W. & Ray Jackendoff. 2005. *Simpler Syntax*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199271092.001.0001.
- Dalrymple, Mary (ed.). 1999. *Semantics and syntax in Lexical Functional Grammar: The resource logic approach* (Language, Speech, and Communication). Cambridge, MA: The MIT Press. DOI: 10.7551/mitpress/6169.001.0001.
- Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III & Annie Zaenen (eds.). 1995. *Formal issues in Lexical-Functional Grammar*. Stanford, CA: CSLI Publications.
- Dalrymple, Mary & Louise Mycock. 2011. The prosody-semantics interface. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '11 conference*, 173–193. Stanford, CA: CSLI Publications.
- Dalrymple, Mary & Irina Nikolaeva. 2011. *Objects and information structure* (Cambridge Studies in Linguistics). Cambridge, UK: Cambridge University Press.
- Falk, Yehuda N. 1983. Constituency, word order, and phrase structure rules. *Linguistic Analysis* 11. 331–360.
- Falk, Yehuda N. 2006. *Subjects and universal grammar: An explanatory theory*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/cbo9780511486265.
- Findlay, Jamie Y. 2016. Mapping theory without argument structure. *Journal of Language Modelling* 4(2). 293–338. DOI: 10.15398/jlm.v4i2.171.
- Halvorsen, Per-Kristian. 1983. Semantics for Lexical-Functional Grammar. *Linguistic Inquiry* 14. 567–615.
- Jackendoff, Ray. 1977.  *$\bar{X}$  syntax: A study of phrase structure* (Linguistic Inquiry Monographs 2). Cambridge, MA: The MIT Press.
- Jackendoff, Ray. 2002. *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.
- Kaplan, Ronald M. 1989. The formal architecture of Lexical-Functional Grammar. *Journal of Information Science and Engineering* 5. 305–322. Revised version published in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 7–27).
- Kaplan, Ronald M. & Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 173–281. Cambridge, MA: The MIT Press. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 29–130).
- Kaplan, Ronald M. & Annie Zaenen. 1989. Long-distance dependencies, constituent structure, and functional uncertainty. In Mark Baltin & Anthony Kroch (eds.), *Alternative conceptions of phrase structure*, 17–42. Chicago: Uni-

*Oleg Belyaev*

- versity of Chicago Press. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 137–165).
- King, Tracy Holloway. 1997. Focus domains and information structure. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '97 conference*, 2–13. Stanford, CA: CSLI Publications.
- Kuhn, Jonas. 2003. *Optimality-Theoretic Syntax – A declarative approach*. Stanford, CA: CSLI Publications.
- Potts, Christopher. 2002. *Comparative economy conditions in natural language syntax*. North American Summer School in Logic, Language, and Information 1, workshop on model-theoretic syntax, Stanford University.
- Pullum, Geoffrey K. & Gerald Gazdar. 1982. Natural languages and context-free languages. *Linguistics and Philosophy*. 471–504. DOI: [10.1007/bf00360802](https://doi.org/10.1007/bf00360802).
- Sadler, Louisa & Rachel Nordlinger. 2004. Relating morphology to syntax. In Louisa Sadler & Andrew Spencer (eds.), *Projecting morphology*, 159–186. Stanford, CA: CSLI Publications.
- Speas, Margaret. 1990. *Phrase structure in natural language*. Dordrecht: Kluwer Academic Publishers. DOI: [10.1007/978-94-009-2045-3](https://doi.org/10.1007/978-94-009-2045-3).
- Wescoat, Michael T. 2002. *On lexical sharing*. Stanford, CA: Stanford University. (Doctoral dissertation).
- Wescoat, Michael T. 2005. English nonsyllabic auxiliary contractions: An analysis in LFG with lexical sharing. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '05 conference*. Stanford, CA: CSLI Publications.
- Woods, William A. 1970. Transition network grammars for natural language analysis. *Communications of the ACM* 13(10). 591–606. DOI: [10.1145/355598.362773](https://doi.org/10.1145/355598.362773).
- Zaenen, Annie & Ronald M. Kaplan. 1995. Formal devices for linguistic generalizations: West Germanic word order in LFG. In Jennifer S. Cole, Georgia M. Green & Jerry L. Morgan (eds.), *Linguistics and computation*, 3–27. Stanford, CA: CSLI Publications. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 215–240).

## Chapter 2

# Core concepts of LFG

Oleg Belyaev

Lomonosov Moscow State University and Institute of Linguistics of the Russian Academy of Sciences

This chapter provides an in-depth coverage of the main features of the LFG framework, focusing mainly on its syntactic representations: c- and f-structure. The makeup of each level is discussed in detail. For c-structure, I describe the version of X-bar theory used in LFG and the status of Lexical Integrity as a core principle of the framework. I discuss the notion of f-structure as a function/set of feature-value pairs that is used in the majority of LFG work; attribute value types and well-formedness conditions on f-structure (Uniqueness, Completeness and Coherence) are covered as well. I also describe the metalanguage for defining f-structures and the mapping from c- to f-structures, and note some linguistically relevant consequences of how this mapping is organized. Three proposed extensions of the standard architecture are also discussed: templates (constructions), minimal c-structure, and lexical sharing.

## 1 Introduction

This chapter provides a detailed survey of the main syntactic levels of LFG, constituent structure (c-structure) and functional structure (f-structure). It complements the more general introduction in Belyaev 2021 [this volume]. In Section 2, I describe the c-structure model used in standard LFG, its understanding of constituency, and the role of X' theory. In Section 3, the notion of f-structure is discussed, including the metalanguage used for describing f-structures and constraints on possible f-structure. In Section 4, I discuss the mapping from c- to f-structure. Finally, in Section 5 I describe recently proposed modifications to the basic architecture of LFG that have not yet been universally accepted, but which may shape the development of this framework in the future.



Oleg Belyaev. 2021. Core concepts of LFG. in Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 21–92. Berlin: Language Science Press. DOI:



*Oleg Belyaev*

## 2 C-structure

The nature of constituent structure (c-structure) in LFG and its main properties are summarized in Belyaev 2021 [this volume]. Briefly, c-structure is a phrase structure tree; constraints on possible trees are usually described via context-free rules as in (1). Other metalanguages are sometimes used as well.

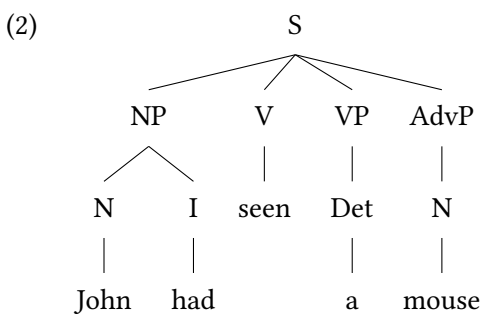
(1)  $S \rightarrow NP\ VP$

Without additional theoretical restrictions, this formalism is too powerful for natural language. In this section, I will focus on two main constraints on c-structure in LFG: X-Bar Theory and lexical integrity.

### 2.1 X-Bar Theory

Every theory of constituency based on phrase structure grammar faces what Everett (2015), in his review of Adger (2013), called “Lyons’ Problem”. Lyons (1968) famously asked what guarantees that NPs are headed by Ns, VPs are headed by Vs, etc., such that rules like  $VP \rightarrow \dots V \dots$  or  $NP \rightarrow \dots N \dots$  are allowed, but rules like  $NP \rightarrow \dots V \dots$  are not.

Indeed, from the point of view of context-free rules, VP and V are atomic symbols that are not related to each other; labeling one of the daughters of NP as N is merely a convention, and nothing in the formalism excludes a hypothetical language with constituent structures like in (2) – “monsters” in Bresnan et al.’s (2016) terms.



Intuitively, there are many things that are wrong with this structure: an I head cannot be the daughter of NP; the VP cannot be headed by, or even immediately

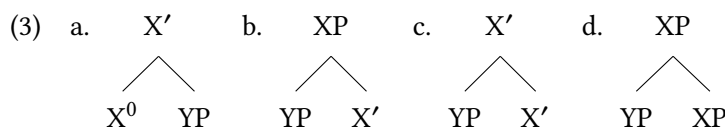
## 2 Core concepts of LFG

dominate, a Det; an AdvP cannot be headed by a noun.<sup>1</sup> The principle that prohibits this is called *ENDOCENTRICITY*; roughly stated, it means that the external distribution of a phrase (e.g. NP) is determined by the category of one and only one of its daughters, the *HEAD*. Disallowing non-endocentric structures requires a theory of constituent structure labels that limits the range of available configurations. To this end, *X-BAR (X') THEORY* has been proposed in mainstream generative grammar (Chomsky 1970; Jackendoff 1977).

X' theory enforces endocentricity by introducing the notion of projection and “bar level” and requiring that each non-maximal projection ( $X^0$  and  $X'$ ;  $X''$ , or XP, is usually assumed to be the maximum level of projection) be dominated by a node belonging to the same category, with the bar level either incremented by one or unchanged. The sisters of c-structure heads (complements, specifiers and adjuncts) have to be maximal projections or non-projecting words (on which see below).

One variant of X' theory has been adopted in LFG from the very early days and continues to be used in most LFG work. An in-depth exposition of X' theory as it is used in LFG, with certain additional theoretical innovations, can be found in Bresnan et al. (2016). The most important features of X' theory as it is practiced in LFG are as follows. First, as in the original formulation, X' theoretical constraints are viewed as constraints on phrase structure *rules*; the later GB view of a kind of universal “X' schema” has not gained acceptance in LFG, primarily because the architecture of the framework is fundamentally based on language-specific rules and does not allow such schemas.

Second, X' theory in LFG allows for the following positions: complement (3a), specifier (3b), X' adjunct (3c) and XP adjunct (3d).<sup>2</sup>



There is some disagreement concerning the possibility of X' adjunction: While most authors accept both kinds of adjunction, Toivonen (2003) only allows XP-

<sup>1</sup>Curiously, each of the features of this illustration *ad absurdum* has a counterpart in real languages: noun phrases do sometimes mark the tense of their clauses, verbs do mark the definiteness of their arguments, and bare nouns (although probably not nouns like ‘mouse’) are used adverbially. But there is broad consensus in theoretical linguistics that such phenomena are more exceptions than rules and should *not* be modeled by allowing the theory of phrase structure to license such configurations.

<sup>2</sup>The order of constituents is only an illustration; X' theory itself does not impose any specific order.

*Oleg Belyaev*

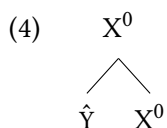
adjunction (and head adjunction, see below) because in her theory only constituents of the same bar level may be adjoined.

The LFG literature also generally allows for multiple complements and specifiers dominated by the same mother node; thus, a sequence of several phrases instead of YP is possible in (3a–c); multiple adjuncts in one position are also usually allowed, even though this creates redundancy since this structure could always be replaced by multiple binary adjunction.

Third, LFG uses the following functional projections: DP for NPs, IP and CP for VPs. Some work also uses additional phrases, such as KP/CaseP for clitic case markers (Broadwell 2008). The number of functional positions is limited compared to mainstream theories, and this is not merely a stipulation: LFG requires all constituency in a given language to be empirically motivated in a way that is more narrow than in frameworks that represent the bulk of syntactic information in phrase structure (such as transformational frameworks). Specifically, heads may only be stipulated if there is actual lexical material that can occupy them; therefore, even the existence of projections such as CP or IP cannot be automatically assumed for all languages. More abstract projections such as TopicP or ForceP are not usually introduced because there are few suitable candidates for the status of heads of these phrases, and little distributional evidence to argue that their specifiers are distinct structural positions.

It turns out, in fact, that the set of functional projections listed above is fully adequate for the overwhelming majority of languages. Moreover, some categories, like DP, are not viewed as universal; some authors, such as Sells (1994) for Japanese and Korean, even limit the number of projections to one ( $X'$ ) instead of the standard two.

Fourth, LFG admits non-projecting words, i.e. lexical items that do not project  $X'$  and XP levels and hence cannot have complements or specifiers; their maximum projection level is 0. The category of non-projecting words is marked as  $X^0$ . Toivonen (2003) develops a detailed theory of non-projecting words. Being maximal projections, they can appear at any non-head  $X'$  theoretic positions (i.e. specifier, complement, or adjunct), but the only dependents that they may have are  $X^0$  adjuncts, which must themselves be non-projecting. Thus, an additional type of adjunction – head adjunction – is introduced into  $X'$  theory, illustrated in (4), where  $X^0$  can also be  $\hat{X}$ , but, crucially,  $\hat{Y}$  cannot be  $Y^0$ , as that would violate the principle that only maximal projections can appear in non-head positions.



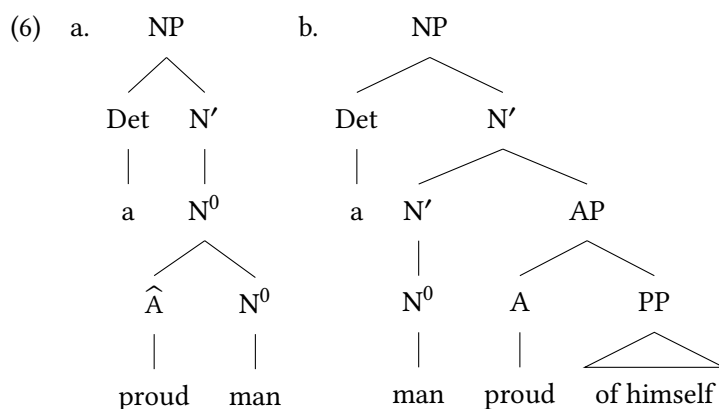
## 2 Core concepts of LFG

The theory of non-projecting words presented in Toivonen (2003) further requires that only same-level projections are adjoined; this effectively prohibits adjoining non-projecting words at  $X'$  or  $XP$  level, as well as any adjunction at  $X'$  level in languages where  $XP$  is the maximal projection. However, these more restrictive principles are not accepted by all authors who use non-projecting words in their analyses: for example, Spencer (2005) analyzes case markers in Hindi as  $\hat{P}$  nodes adjoined to  $NP$ .  $X'$  adjunction also remains quite common in LFG analyses.

Sadler & Arnold (1994) use non-projecting words to account for the behaviour of English prenominal adjectives, which cannot have phrasal complements if they are prenominal; consider the contrast between (5a) and (5b), while (5c) is ungrammatical.

- (5) a. a **proud** man  
 b. a man [**proud** of himself]  
 c. \* a [**proud** of himself] man

Sadler and Arnold argue that this contrast is due to the fact that prenominal adjectives in English are non-projecting words with the category  $\hat{A}$  that are adjoined to  $N^0$ , while postnominal adjectives form  $AP$  and can therefore have complements. Thus the structure of (5a) is (6a), while the structure of (5b) is (6b).

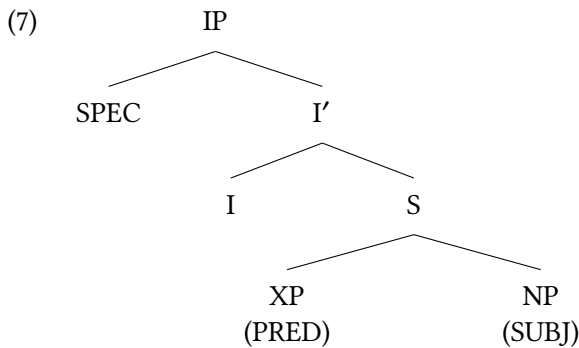


Finally,  $X'$  theoretic principles are not viewed as fully universal in LFG. The most prominent exception is the exocentric category  $S$ .<sup>3</sup> This category does not have a “head” in the normal sense: it can be “headed” by a verb, but also by an

<sup>3</sup>Bresnan et al. (2016: 112ff.) present the category  $S$  and non-projecting words as effectively the *only* exceptions from standard principles of  $X'$  theory. This, however, is a theoretical idealization insofar as it applies to actual LFG analyses, which routinely make use of *ad hoc* categories

*Oleg Belyaev*

adjective or another nonverbal predicate; this is why the term S is used instead of, for example, VP. The category S is most extensively used in nonconfigurational languages (see **chapters/Cstr**), but this is not its exclusive role. Many languages have a fairly configurational structure overall but allow predicates of various categories to be embedded under a general “predicative marker”, which sits in the I or C node. For example, Kroeger (1993: 119) proposes the phrase structure in (7) for Tagalog. The SPEC position can be optionally occupied by fronted constituents of several types (such as topics); the I node is occupied by an auxiliary or the finite verb; the predicate XP can be a VP in verbal sentences, but can also be AP or NP if the predicate is nonverbal. Hence, the structure is indeed non-endocentric, and the use of the label S is justified.



Since c-structure is not the only level of representation in LFG and models only a subset of syntactic phenomena (word order, embedding), X' theory does not do much by itself to limit the range of possible languages. Unlike frameworks such as GB, for which the theory was originally devised, X' positions are not inherently or uniquely associated with specific syntactic or semantic functions – as a result, X' theory, understood purely in terms of c-structure, is little else than a system of labeling nodes which allows us to generalize endocentricity at constituent structure level. In order to make it more meaningful, it should be augmented by a set of principles that determine the mapping of X' positions to f-structure – such a system has been developed in LFG, and will be described in Section 4.3.

---

such as CL, CCL (for “clitic”, “clitic cluster”) in Bögel et al. (2010) and Lowe (2011). Such minor innovations do not seem to influence the overall theory in any meaningful way, since they deal with exceptional cases such as second-position clitics or language-specific, idiosyncratic linear order distributions. It is also conceivable that many of them could be converted to analyses that conform to X' theoretic principles; for example, CCL could be treated as a phrase consisting of multiple  $\hat{D}$  head adjunction (if the clitics are pronominal).



## 2.2 Lexical Integrity

As its name implies, Lexical Functional Grammar was originally conceived as a lexicalist framework, a term that has several meanings. In the most general sense, lexicalism implies that the features of individual syntactic elements (morphemes and wordforms) as well as their subcategorization frames are determined in the lexicon, and cannot be modified in the syntax (such as by promoting the direct object in a passive construction). Lexicalism in this sense requires no additional stipulation and is enforced by the LFG architecture itself: there are no transformations or other means to change the c-structure or f-structure features; syntax can only multiply define lexical features, but cannot override them.<sup>4</sup>

LFG is also lexicalist in another sense: it subscribes to the idea that the building blocks of syntax are not roots or affixes, but individual words that are constructed from different blocks and according to different rules than syntactic constituents.<sup>5</sup> Thus, the distinction between morphology and syntax in LFG is viewed as fundamental, which is against the views of many recent approaches, both formal (Bruening 2018) and typological (Haspelmath 2011).

This understanding of lexicalism is more formally termed *LEXICAL INTEGRITY* and has been given two formulations in LFG (8)–(9).

- (8) Words are built out of different structural elements and by different principles of composition than syntactic phrases. (Bresnan & Mchombo 1995: 181)
- (9) Morphologically complete words are leaves of the c[onstituent]-structure tree and each leaf corresponds to one and only one c[onstituent]-structure node. (Bresnan et al. 2016: 92)

The definition in (8) is rather broad and can be compatible with several different understandings of the morphology–syntax interface, as long as the border between the two levels is maintained in some way. The second definition (9) is more specific and is only compatible with one view of the interaction between

<sup>4</sup>Kaplan & Wedekind (1993) introduced the restriction operator:  $f \setminus_A$  denotes the f-structure  $f$  with the attribute  $A$  and its value removed. As suggested by an anonymous reviewer, this violates lexical integrity in the weak sense, because here the syntax effectively accesses an f-structure constructed otherwise (possibly by means of morphology) to retrieve some of its information. This operator is not widely employed but was used in several LFG analyses, notably in Asudeh (2012) and Falk (2010).

<sup>5</sup>These two understandings of lexicalism are sometimes conflated, but they are actually independent: A framework may be lexicalist in the former sense, but consider the distinction between words and syntactic phrases to be ephemeral.

*Oleg Belyaev*

morphology and syntax. For example, lexical sharing, which allows one word to correspond to two  $X^0$  nodes (discussed in Section 5.2.2), is compatible with (8) but not with (9).

Interestingly, despite the rather strict definition in (9), much work in LFG uses the concept of “sublexical nodes”, like in the rule for Greenlandic nouns in (10), from Bresnan et al. (2016: 368). This is formally incompatible with (9) because the preterminal nodes correspond to morphemes, not morphologically complete nodes.

$$(10) \quad N \rightarrow N_{\text{stem}} \quad N_{\text{aff}}$$

In practice, such analyses are rather harmless because in their predictions they are equivalent to analyses that strictly adhere to (9): since all morphology is sublexical, the position of individual affixes cannot have any syntactic relevance, as opposed to approaches like Distributed Morphology (Halle & Marantz 1993), where morphological features often occupy higher-level functional projections that can scope over syntactic phrases. However, the use of sublexical representations does raise the issue of how the individual contribution of morphemes to f-structure should be represented – standard LFG does not provide such a way, because words are viewed as complete, unsegmented bundles of morphosyntactic information. These issues are discussed in detail in **chapters/Morphology**.

## 3 F-structure

### 3.1 The notion of f-structure

As described in Belyaev 2021 [this volume], c-structure in LFG is complemented by an additional level of representation called F-STRUCTURE. F-structure is an attribute-value structure that includes information on valency, grammatical functions, and the features of clauses and their syntactic arguments. An f-structure for the English sentence *John has seen David* is given in (11).

$$(11) \quad \left[ \begin{array}{ll} \text{PRED} & \text{'SEE'} \langle (f \text{ SUBJ}) (f \text{ OBJ}) \rangle \\ \text{TENSE} & \text{PRS} \\ \text{ASPECT} & \text{PERF} \\ \text{SUBJ} & \left[ \begin{array}{ll} \text{PRED} & \text{'JOHN'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{array} \right] \\ \text{OBJ} & \left[ \begin{array}{ll} \text{PRED} & \text{'DAVID'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{array} \right] \end{array} \right]$$

## 2 Core concepts of LFG

F-structure is usually thought of as a set of attribute-value pairs, or a function that maps attribute names to their values. This understanding of f-structure has important implications for the architecture of LFG. Specifically, it implies that f-structures are solely and uniquely defined by their set of attribute-value pairs; there is no type system as in Carpenter (1992) or HPSG (Pollard & Sag 1994). Therefore, there is no such thing as two different f-structures having the same set of attributes and values; the notion of an empty f-structure is also problematic, because all empty structures are equivalent to each other.<sup>6</sup> This notion of identity is somewhat mitigated by the uniqueness of PRED values (Section 3.3.4), which ensures that any two independently introduced, semantically interpreted f-structures are formally distinct, even if they have the same lexical predicate and the same set of morphosyntactic features. However, not all f-structures have PRED values; thus, for instance, all expletive subjects of the same form are described by the same f-structure, regardless of the clauses in which they occur. For example, all bundles of agreement features (AGR) with the same set of values are identical to each other. One AGR bundle may be required to be identical to another via agreement sharing (Haug & Nikitina 2015) in the f-description (using an equation such as  $(\uparrow \text{AGR}) = (\uparrow \text{SUBJ AGR})$ ), but it will also be identical to all other such bundles elsewhere in the same sentence, if they occur. Counterintuitive though such results may seem, it is not clear whether they can lead to any undesirable effects in practice. The notion of identity of f-structures is important for understanding the concept of STRUCTURE SHARING (Section 3.3.3).

## 3.2 The metalanguage

### 3.2.1 Defining equations

The standard notation for describing f-structures are DEFINING EQUATIONS, which utilize the idea that f-structures are functions. For example, the value of the attribute TENSE of f-structure  $f$  in (11) can be defined by the equation  $(f \text{ TENSE}) = \text{PRS}$ . It is possible to use nested function applications; thus, since  $(f \text{ SUBJ}) = g$ ,  $((f \text{ SUBJ}) \text{ PERS})$  is equivalent to  $(g \text{ PERS})$  and has the value 3. By convention, function application is left associative, thus the parentheses can be omitted and the equation written as  $(f \text{ SUBJ PERS}) = 3$ .

Defining equations are grouped into F-DESCRIPTIONS. An f-description describes the *minimal* f-structure that satisfies all the equations included in the description.

---

<sup>6</sup>Observe that the standard LFG notation does not even have a way to specify empty f-structures, on the tacit assumption that every non-vacuous f-structure would have at least one attribute. However, the notion could be useful e.g. for expletive arguments that are not specified for any morphosyntactic features such as PERS or CASE but simply appear to satisfy Completeness.

Oleg Belyaev

The default relation between equations forming an *f*-description is conjunction, but disjunction is also possible; for example,  $\{(\uparrow \text{SUBJ PERS})=1 \mid (\uparrow \text{SUBJ PERS})=2\}$  means that the subject is defined as being either 1st or 2nd person.<sup>7</sup> For more examples and discussion of defining equations, see Belyaev 2021 [this volume].

### 3.2.2 Constraining equations

The *f*-structure equations described above are all evaluated to construct the minimal complete and coherent *f*-structure that satisfies all of them together (if such an *f*-structure exists). In this sense, they are “constructive”, or *DEFINING*: informally, a defining equation introduces a feature value, regardless of whether it is the only such equation or the same value is defined elsewhere.

But sometimes it is necessary to check the value of a feature without actually assigning it. For example, a matrix verb might require its complement to have a specific mood value, such as subjunctive. A defining equation like  $(\uparrow \text{COMP MOOD})=\text{SBJV}$  also licenses a complement that lacks a *MOOD* feature altogether (e.g., it is non-finite), which probably leads to an incorrect prediction (unless additional constraints block the use of such forms in this context).

Defining equations also provide no way to capture purely negative requirements, i.e. to ensure that a feature *does not* have a specific value. Clearly, this is not equivalent to the disjunction of other possible values of the feature, since, first, absence of the feature also satisfies the negative condition; second, the disjunction would freely assign any feature value except for the disallowed one, which is definitely not what a negative constraint should do.

The need for such constraints is accounted for in LFG by allowing a special class of equations, *CONSTRAINING EQUATIONS*. These equations are special in that they do not participate in constructing the *f*-structure of the sentence. In contrast, they are only evaluated once the minimal *f*-structure satisfying all defining equations has been constructed. Then, violation of a constraining equation leads to ungrammaticality.

The simplest type of constraining equations involve equality relations; these are annotated in the same way as defining equations, but with a subscript *c*, e.g.:  $(fa) =_c x$ . To illustrate how constraining equations work, consider the following *f*-descriptions and their corresponding *f*-structures:

$$(12) \quad \begin{array}{ll} \text{a.} & \begin{array}{l} (f \text{ A}) = x \\ (f \text{ B}) = y \\ (f \text{ A}) =_c x \end{array} \end{array} \rightarrow f \begin{bmatrix} \text{A} & x \\ \text{B} & y \end{bmatrix} \quad \text{(constraining equation satisfied)}$$

<sup>7</sup>Disjunction can be represented by either a vertical line (|) or a logical disjunction sign (v); both notations are found in the literature.

## 2 Core concepts of LFG

$$\begin{array}{lcl} \text{b.} & (f \text{ B}) = \text{Y} & \nleftrightarrow f[\text{B Y}] \quad (\text{constraining equation not satisfied}) \\ & (f \text{ A}) =_c \text{X} & \end{array}$$

In (12a), the constraining equation is satisfied because the feature value is defined elsewhere. By contrast, in (12b) the constraining equation is not satisfied because A has no value, and a value cannot be assigned by a constraining equation.

Note that constraining equations serve as a good illustration of the LFG principle of separation between description and the object being described. Just as multiple feature definitions are not represented in the *f*-structure, there is also no trace of constraining equations having been “checked” in (12a). The only thing a constraining equation does is to put constraints on permissible structures; it does not contribute to the structures themselves.

The other two types of constraining equations are **EXISTENTIAL** and **NEGATIVE** constraints. Existential equations check that a feature has *any* value rather than testing for a specific value. They are written as simple function applications:  $(fa)$  means that the *f*-structure *f* must have the feature *a* with any value; the absence of an equality statement indicates that we are dealing with an existential constraint. Negative constraints check that a feature does not have a given value  $((fa) \neq x$ ; this is compatible with the feature having no value) or has no value  $(\neg(fa))$ ; this is called a negative existential constraint).

Constraining equations are also implicitly introduced by **CONDITIONAL STATEMENTS** of the form  $X \Rightarrow Y$ . These are, by definition (Bresnan et al. 2016: 61; Dalrymple et al. 2019: 168), equivalent to a disjunction:  $\neg A \vee (A_c \wedge B)$ .

**OFF-PATH CONSTRAINTS** are conceptually similar to conditional statements in that they are used to restrict function application to apply only to *f*-structures satisfying additional conditions on their features. For example,

$$(f \begin{array}{c} \text{A} \\ (\rightarrow \text{B}) =_c \text{y} \end{array} \text{c}) = x$$

means that the value *x* is only assigned to the feature *c* of the *f*-structure  $(f \text{ A})$  if  $(f \text{ A})$  has an attribute *B* with the value *y*. If only constraining equations are used in such statements (as assumed in some of the literature), they could all in principle be rewritten as conditional statements (provided that local names are used), but the notation is more cumbersome. This is indeed assumed in some LFG literature, and perhaps most prominently in the XLE implementation (Crouch et al. 2008), where defining equations cannot be used in off-path constraints (see Patejuk & Przepiórkowski 2014: 7 for a discussion). However, in spite of their name, the theoretical literature (Bresnan et al. 2016: 65, fn. 26; Dalrymple et al. 2019: 230) unanimously suggests that off-path constraints can be constructive,

*Oleg Belyaev*

and this feature is used in some LFG analyses.<sup>8</sup> Off-path constraints are especially important for Functional Uncertainty expressions (Section 3.2.3), where a path may be a regular expression with many elements and the direct use of conditional statements is impractical.

It is clear from the discussion above that while the concept of constraining equations appears rather simple, it actually introduces some additional complexity into the system. Instead of just evaluating an *f*-description that consists of a set of defining equations, the resolution of a valid *f*-structure for a sentence must proceed through two steps: (a) evaluation of defining equations; (b) evaluation of constraining equations. The notion of constraining equations has also raised concerns about the metatheoretical status of LFG grammars; in particular, Pullum (2013) and Blackburn & Gardent (1995) have argued that constraining equations introduce a degree of procedurality into the framework, which is incompatible with the notion of model-theoretic syntax. However, the specific implications of this procedurality have never been systematically studied. It is clear that many, perhaps most, grammars that use constraining equations could be rewritten without them, but with more notational complexity: for example, by requiring every *f*-structure to have certain attributes, introducing “empty” attribute values (i.e. treating “no value” as one of the values for atomic features), and so on. Thus the issue might, in the end, be more of notation rather than substance, as suggested, in fact, in the conclusion to Blackburn & Gardent (1995).

### 3.2.3 Functional uncertainty

The basic LFG architecture outlined in the preceding sections is adequate to handle most phenomena that are relevant to the local structure of clauses and noun phrases, such as argument selection and realization, modification, and word order. However, it is missing a component that could handle unbounded dependencies of any kind, i.e. those dependencies between elements of a sentence that are not tied to any specific structural position. For example, consider the behaviour of “cyclic” extraction from complement clauses. This process is in principle unbounded: an interrogative might be extracted from the matrix clause (13a), from the complement clause (13b), from the complement of the complement (13c), etc.

- (13) a. **Who** does John like \_?  
       b. **Who** does John think Mary likes \_?  
       c. **Who** does John believe David thinks Mary likes \_?

---

<sup>8</sup>I am thankful to an anonymous reviewer for drawing my attention to this fact.

## 2 Core concepts of LFG

For (13a), one might write an f-structure equation annotating the extracted NP node such as  $(\uparrow \text{OBJ}) = \downarrow$ , and augment it with a disjunction for each other available grammatical function – which, by itself, is not very elegant, but seems to adequately account for the facts. To capture (13b), another set of equations must be added to the disjunction, this time with COMP before GF:  $(\uparrow \text{COMP OBJ}) = \downarrow$ , etc. This already seems like a rather artificial solution, but when (13c) is considered, yet another disjunction is required:  $(\uparrow \text{COMP COMP OBJ}) = \downarrow$ , etc. Clearly, the sequence of COMP's can be arbitrarily large (if memory constraints and other extralinguistic considerations are not taken into account), and any grammatical framework must account for such boundless iteration. LFG, in its basic form described above, clearly cannot do so.

Intuitively, what is required is to allow generalizing over sets of functional equations, specifically, introducing disjunction to allow selecting different GFs, and arbitrary iteration of COMP. This is achieved by the notion of FUNCTIONAL UNCERTAINTY, introduced to LFG in Kaplan & Zaenen (1989b). In a nutshell, functional uncertainty extends the LFG notion of function application by allowing function names –  $x$  in a statement like  $(fx)$  – to be regular expressions. Thus, a single f-structure equation may correspond to a (possibly infinite) set of statements. More formally, functional uncertainty defines function application as in (14).

- (14)  $(f \alpha) = v$  holds if and only if  $f$  is an f-structure,  $\alpha$  is a set of strings, and for some  $s$  in the set of strings  $\alpha$ ,  $(f s) = v$ .

Thus, the distribution in (13) can be captured by a single equation, such as in the following rule for extracted interrogatives:

- (15)  $\text{CP} \longrightarrow \begin{array}{cc} \text{NP} & \text{C}' \\ (\uparrow \text{DIS}) = \downarrow & \uparrow = \downarrow \\ (\uparrow \text{COMP}^* \{ \text{OBJ} \mid \text{OBJ}_{\theta} \mid \text{OBL}_{\theta} \}) = \downarrow & \end{array}$

The disjunction in the NP annotation is typically abbreviated as GF, which stands for “any grammatical function” – but which GFs exactly can appear in a given position is construction-specific; for example, adjuncts may or may not be included in the list of GFs. In general, so-called “island constraints” are typically captured in LFG as constraints on paths in functional uncertainty equations (Kaplan & Zaenen 1989b). This correctly predicts that what counts as an “island” varies across languages and across different constructions within the same language.

Oleg Belyaev

### 3.2.4 Inside-out function application and functional uncertainty

Standard function application in LFG is “outside-in”: an expression  $(fa)$  refers to a feature that belongs to the f-structure  $f$  or at any deeper level of embedding. This presupposes a “top-down” style of describing and constraining f-structures. However, it may sometimes be useful to describe constraints on the *external* distribution of an f-structure: for instance, limit the range of attributes it may occupy, or define some features of “sister” f-structures, i.e. f-structures that occupy different attributes in the containing f-structure (e.g., SUBJ constraining attributes of OBJ). For this, LFG uses an additional mechanism called **INSIDE-OUT EXPRESSIONS**. Inside-out expressions use the same parenthetical notation as ordinary LFG notation, but the f-structure now acts as an argument rather than as a function. Formally, inside-out expressions are defined as follows:

- (16)  $(a f) = g$  holds if and only if  $g$  is an f-structure,  $a$  is a symbol, and the pair  $\langle a, f \rangle \in g$ .

Informally, this definition means that  $(a f)$  refers to an f-structure  $g$  (or set of f-structures) whose attribute  $a$  has  $f$  as its value. For example, in (17),  $(A g) = f$  holds because  $(f A) = g$  is satisfied.

$$(17) \quad f \left[ A \quad g \left[ B \quad x \right] \right]$$

Functional uncertainty can also be extended to cover inside-out expressions by replacing  $a$  in the definition above by a regular expression  $\alpha$ . The formal definition is as follows:

- (18)  $(\alpha f) \equiv g$  if and only if  $g$  is an f-structure,  $\alpha$  is a set of strings, and for some  $s$  in the set of strings  $\alpha$ ,  $(s f) = g$ .

Inside-out function application is by its nature a rather limited formal device compared to “outside-in” function application. It is mainly used either to constrain the grammatical functions that an f-structure may occupy, or to constrain the features of this f-structure. Importantly, it cannot actually be used as the main mechanism of constructing f-structures. For example, one may formulate a defining equation such as  $((A f)A) = f$  to force  $f$  to appear in grammatical function A. But this definition will produce an “orphaned” f-structure which can only be integrated with other f-structures by additional “outside-in” statements, which, in turn, make such an inside-out statement redundant.

Which grammatical phenomena is inside-out functional uncertainty used to model? Perhaps the simplest is the restriction of certain grammatical forms to



## 2 Core concepts of LFG

certain syntactic positions. For example, if nominative marking in a given language is always associated with the grammatical function SUBJ, one may avoid referring to a feature CASE, instead adding (SUBJ  $\uparrow$ ) to the lexical entries of all nominative nouns. This correctly ensures that nominative nouns are only used in those positions which the grammar defines as being associated with subjects. A prominent example of this approach to case is described in Nordlinger (1998).

Another phenomenon where inside-out function application plays a role is agreement on modifiers. For example, Russian adjectives agree in gender and number with their heads. In standard LFG terms, this means that they are lexically annotated to co-describe the features CASE and NUM of the f-structures whose ADJUNCT position they occupy. An adjective like *krasnaja* ‘red’ (fem. sg.) might have the following lexical entry:<sup>9</sup>

- (19) *krasnaja*    Adj    ( $\uparrow$  PRED) = ‘RED’  
                                   ((ADJ  $\in$   $\uparrow$ ) NUM) = SG  
                                   ((ADJ  $\in$   $\uparrow$ ) GEND) = FEM

A somewhat more exotic phenomenon that inside-out functional uncertainty succeeds at capturing is “case-stacking” in Australian languages, where NP-internal dependents are marked not only with the case that indicates their position within this NP, but also for the case that indicates the position of this NP at a higher level. Nordlinger (1998) develops a theory called Constructive Case<sup>10</sup> to account for this behaviour; “stacked” cases are treated as denoting the case values of the f-structures that contain the noun as their complement, via the mechanism of inside-out functional uncertainty.

### 3.2.5 Local names

F-structures in annotated rules are typically referred to relative to the nodes in the phrase structure rule, i.e. using paths that begin in the metavariables  $\uparrow$  or  $\downarrow$ . This is sufficient if these paths are uniquely and unambiguously resolved; the relevant reference may just be repeated in all equations that use it. But in some cases functional annotations do not uniquely identify f-structures that should be referred to. This most frequently occurs when functional uncertainty is involved

<sup>9</sup>The set membership symbol  $\in$  may be used in inside-out statements just as well as it can be used in outside-in statements. ( $A \in f$ ) =  $g$  entails that ( $g A \in$ ) =  $f$ , which is notationally equivalent to  $f \in (g A)$ .

<sup>10</sup>“Constructive” is, in fact, somewhat of a misnomer: as shown above, inside-out statements cannot actually “construct” anything, but can only test for feature values of f-structures that have already been constructed.

*Oleg Belyaev*

(described in Section 3.2.3), i.e. when paths are regular expressions that can resolve to different f-structures. Another example is when the same set of rules can apply to different f-structures, either in free variation or subject to additional conditions. Consider a hypothetical language where verbal agreement morphology can alternatively define the person and number features of the subject or direct object.<sup>11</sup> In this case, it is of course possible to introduce disjunction of two sets of equations, as in (20), but this clearly misses the crucial generalization that the same features are defined in both disjuncts.

$$(20) \quad \left\{ \begin{array}{l} (\uparrow \text{ SUBJ PERS}) = 1 \\ (\uparrow \text{ SUBJ NUM}) = \text{SG} \\ | (\uparrow \text{ OBJ PERS}) = 1 \\ (\uparrow \text{ OBJ NUM}) = \text{SG} \end{array} \right\}$$

A more economical way to formulate this constraint is to introduce a temporary label for the f-structure involved – a *LOCAL NAME* – and then refer to this name in the two equations assigning person and number features. Normal names in LFG, by convention, are written with an initial % and assigned using the standard equation operators, as in (21):

$$(21) \quad \left\{ \begin{array}{l} \% \text{AGR} = (\uparrow \text{ SUBJ}) \mid \% \text{AGR} = (\uparrow \text{ OBJ}) \\ (\% \text{AGR PERS}) = 1 \\ (\% \text{AGR NUM}) = \text{SG} \end{array} \right\}$$

While local names are not very frequent in LFG analyses, their use is essential for some phenomena where there is a need to consistently refer to an f-structure whose identity is not uniquely deducible from its path (set members, functional uncertainty, etc.).

### 3.2.6 F-precedence

The basic architecture of LFG is devised to be modular, such that different linguistic phenomena are accounted for at separate levels. In the interaction between c- and f-structure, c-structure is exclusively concerned with linear order and hierarchical embedding, while f-structures do not reflect linear order or constituent structure in any way. Therefore, linear order is relevant for most morphosyntactic constraints only in a limited way, insofar as it distinguishes between different c- to f-structure mappings (such as, for example, in English, where Spec,IP

---

<sup>11</sup>This example is actually not so hypothetical: such an analysis of Dargwa is proposed in Belyaev (2013), where person-number agreement can be associated with either subject or object, and the choice is then “filtered” using a set of OT constraints.

## 2 Core concepts of LFG

is mapped to subject and precedes the verb and Comp,VP). Without extensions to the standard LFG notation, there is no way to state a constraint like “the verb agrees in person and number with whatever NP stands to its left”, because agreement features are the domain of f-structure, and functional equations can only refer to f-structure functions, not linear or constituent-based positions.

However, in certain cases linear order does seem to play a role in determining constraints on syntactic relations. A well-known example is the availability of discourse anaphora between adverbial clauses and main clauses: If the antecedent precedes the pronoun, coreference is possible regardless of which clauses the two are located in (22), while cataphora (backwards anaphora) is only possible if the cataphor stands in the subordinate clause (23).

- (22) a. [ When John<sub>i</sub> came ], I saw him<sub>i</sub>.  
 b. I saw John<sub>i</sub> [ when he<sub>i</sub> came ].
- (23) a. [ When he<sub>i</sub> came ], I saw John<sub>i</sub>.  
 b. \* I saw him<sub>i</sub> [ when John<sub>i</sub> came ].

Such behaviour has been generalized since Langacker (1969) as “precede-and-command”.<sup>12</sup> Coindexation is possible if at least one of the following is true: the antecedent c-commands<sup>13</sup> the pronoun; the antecedent precedes the pronoun.

Similar constraints operate in other languages. For example, Mohanan (1982) argues that in Malayalam, pronouns *must* follow their antecedents. In LFG, such constraints can be captured using the relation of F-PRECEDENCE (Kaplan & Zaenen 1989a), which is a way of introducing linear order constraints in f-structure using the inverse projection  $\phi^{-1}$ , which maps f-structures to the corresponding c-structure nodes.

- (24)  $f$  *f-precedes*  $g$  ( $f <_f g$ ) if and only if for all  $n_1 \in \phi^{-1}(f)$  and for all  $n_2 \in \phi^{-1}(g)$ ,  $n_1$  c-precedes  $n_2$ .

The formal definition in (24)<sup>14</sup> essentially means that an f-structure  $f_1$  f-precedes  $f_2$  iff all c-structure constituents that map to  $f_1$  linearly precedes the constituent

<sup>12</sup>The relevance of linear order has been hotly contested in the literature on anaphora, especially in mainstream transformational grammar; for a recent take on precede-and-command, see Bruening (2014). This is not relevant for our discussion, though, as within LFG no one ever argued against linear-order constraints on anaphora.

<sup>13</sup>In LFG, c-command is replaced by outranking on the grammatical function hierarchy: see [chapters/Anaphora](#).

<sup>14</sup>C-precedence requires that all daughter nodes of a node precede all daughter nodes of another node – essentially a linear precedence relation for c-structure constituents.

*Oleg Belyaev*

that maps to  $f_2$ . Given this definition, anaphoric constraints such as precede-and-command may be formulated as the requirement that the pronoun's antecedent f-precede the pronoun.

Note that f-precedence is a rather straightforward relation if an f-structure corresponds to a single constituent. In more complex situations, such as when discontinuous constituents are involved, or one of the elements does not have a c-structure exponent, its application is not so intuitive. In particular, in the latter case, null elements f-precede and are f-preceded by all other elements in the sentence, because one of the sets  $n_1, n_2$  is empty. This property of f-precedence is used to analyze the behaviour of null anaphora in languages like Malayalam (Mohanam 1982) or Japanese (Kameyama 1985), where null pronouns behave differently from full pronouns. For such languages, the definition in (24), combined with the generalization in the preceding paragraph, correctly predicts that linear order does not influence the anaphoric requirements of null pronouns (Dalrymple et al. 2019: 257).

An alternative definition of f-precedence, that leads to a different treatment of null pronouns, is proposed in Bresnan et al. (2016: 213):

- (25)  $f$  f-precedes  $g$  if and only if the rightmost node in  $\phi^{-1}(f)$  precedes the rightmost node in  $\phi^{-1}(g)$ .

Under this definition, null pronouns in fact do not f-precede and are not f-preceded by any constituent, because their inverse projections lack a rightmost node. To capture the data of Japanese or Malayalam using this definition, a different, negative formulation of the precedence binding constraint should be used: “The domain of a binder *excludes* any pronominal that f-precedes it” (Bresnan et al. 2016: 213, emphasis mine), i.e. the pronoun *must not* f-precede its antecedent. For more information on f-precedence and linear order constraints on anaphora in general, see **chapters/Anaphora**.

Thus, the use of inverse projection does allow a degree of influence of linear order on syntactic constraints, in a limited way (as intended): linear order may serve as an additional constraint on relations formulated in f-structure terms, but does not serve as the only or as the main factor determining these relations.

### 3.3 Attribute value types

#### 3.3.1 General remarks

The system of attribute values in the core LFG architecture is very straightforward. There are only three types of values: atomic values, semantic forms and other f-structures (of which sets are a special instance).

## 2 Core concepts of LFG

The simplicity of this system follows from the fact that, as mentioned above, LFG has no type system for f-structures. This means that the list of potential attributes and their values for any given f-structure is defined only by annotated phrase structure rules and lexical entries. Thus, there is nothing in the formal architecture or in any part of an LFG grammar that would prohibit a “clausal” f-structure to have the feature *CASE* or a “nominal” f-structure to have the feature *TENSE*; such constraints are only implicit in the way these f-structures are constructed and mapped from c-structure nodes.

Similarly, the attributes themselves are not by default associated with any specific value type: LFG grammars by themselves contain no stipulation of possible attributes and the values they may take. Only grammatical function values are required to be f-structures, and *PRED* values to be semantic forms due to Completeness and Coherence (see Section 3.4). Nothing prevents the value for *CASE* or *PERS* to be an f-structure rather than an atomic value; in fact, the former option has been used in analyses such as Dalrymple & Kaplan (2000).

This simplicity of the type system may be viewed as an advantage, as it simplifies the LFG metalanguage without introducing unnecessary redundancy (see Asudeh & Toivonen 2006: 412ff. for a criticism of the Minimalist feature system). There are few problems that a more complex type system would solve, as the architecture of a well-defined grammar typically prevents f-structures from being assigned incorrect attribute values. Still, sometimes it is necessary to check that an f-structure belongs to a given type – for example, whether it is nominal or clausal. LFG provides several ways to do so: one might directly check the category of the corresponding c-structure node using an inverse projection (Section 4.1), or check for certain characteristic attributes (such as *CASE* for nominals or *TENSE* for finite clauses) using constraining equations. The latter method, however, is error-prone, as the grammar writer has to ensure that all relevant f-structures have these attributes. This issue can be partly remedied using templates (Asudeh et al. 2013), but templates are an optional, purely notational device; care must be taken that templates are used consistently.

Another solution has been implemented in XLE, which allows the grammar writer to optionally use *FEATURE DECLARATIONS* to describe the restrictions on feature values (Crouch & King 2008). This is a robust system which, if employed properly, can provide grammars with a higher degree of generalization while also decreasing the number of accidental errors in feature descriptions. Unfortunately, it is virtually unknown in the LFG theoretical literature, being meant as an engineering solution rather than a theoretical proposal and limited to computational work that uses XLE (see Forst & King 2021 [this volume] for more detail).

*Oleg Belyaev*

### 3.3.2 Atomic values

The simplest type of attribute value is an atomic value: essentially a token that represents a given value of a grammatical feature (e.g. ACC for the feature CASE, PRESENT for the feature TENSE, etc.). There is no single agreed-upon set of “standard” features and the valid values they might take: in principle, it is the task of the grammar writer or analyst to determine the set of features required to describe a particular language.

In current LFG practice, there is, however, a set of informal conventions on the general inventory of atomic features. These fall into two types. The first type are morphosyntactic features of the same kind as those standardly used in typology and descriptive grammars: features such as CASE, TENSE, ASP, PERS, etc. An overview of the use of features in syntactic and morphological description can be found in Corbett (2012).

The second type are more technical features that are specific to the LFG understanding of specific syntactic phenomena. For example, Dalrymple (2001: 396ff.) uses the feature LDD (for Long-Distance Dependency) to mark whether an  $f$ -structure is available for extraction. If  $(f \text{ LDD}) = -$ , the  $f$ -structure  $f$  cannot be in the path that specifies a long-distance dependency. This feature is checked by an off-path constraint (see Section 3.2.2). These and similar constraints are discussed in more detail in **chapters/LDDs**.

Similarly, features such as PRONTYPE or NUCLEAR are used in Dalrymple (1993); Bresnan et al. (2016) to distinguish between different kinds of pronouns to account for the differences in binding constraints. See **chapters/Anaphora** for more detail.

In spite of the theoretical significance and cross-linguistic ubiquity of such features as LDD and PRONTYPE, it is generally assumed that they are also not universal and not part of an innate grammatical blueprint (although, to my knowledge, this question has never been explicitly discussed in the literature). Thus, while Bresnan et al.’s (2016) approach to anaphora relies on grammar-wide constraints and distinguishes pronouns via their features, Dalrymple (1993) rather assumes that all binding constraints are lexically specified by the pronouns themselves. The latter point of view is supported by the cross-linguistic diversity of binding domains. It might be that both approaches are valid, but the efficiency of each depends on the language in question. Hence, like in many other domains, LFG as a framework is agnostic as to whether cross-linguistic similarities are due to innate, universal constraints or are a result of independent, functionally motivated convergence of grammars in the course of their evolution. Particular analyses

## 2 Core concepts of LFG

can strike a balance between these two factors that explain cross-linguistic similarities.

## 3.3.3 F-structure

As seen in (11), f-structures can themselves serve as attribute values. F-structures are predominantly values of grammatical functions such as SUBJ, OBJ, etc., and discourse functions such as TOPIC and FOCUS (see **chapters/GFs**). F-structures are sometimes also used to represent “compound” attribute values; for example, agreement features are sometimes represented as the “bundle” AGR in (26), and PRED values can be viewed as composite (Section 3.3.4).

$$(26) \begin{bmatrix} \text{PRED} & \text{'HOUSE'} \\ \text{AGR} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix}$$

Just as different atomic-valued attributes can have identical values, one f-structure can also serve as a value for several attributes. This phenomenon is called **STRUCTURE SHARING** and is the closest LFG counterpart to the notion of “movement” in transformational frameworks; it is discussed in more detail in Belyaev 2021 [this volume]. This configuration can be visually represented in two ways: either the f-structure is fully spelt out in every occurrence (27a), or only once – then the other occurrences are connected by lines (27b) or coindexed (27c).

$$(27) \begin{array}{ll} \text{a.} & \begin{bmatrix} \text{ATTR1} & \begin{bmatrix} \text{A1} & \text{v1} \\ \text{A2} & \text{v2} \end{bmatrix} \\ \text{ATTR2} & \begin{bmatrix} \text{A1} & \text{v1} \\ \text{A2} & \text{v2} \end{bmatrix} \end{bmatrix} \\ \text{b.} & \begin{bmatrix} \text{ATTR1} & \begin{bmatrix} \text{A1} & \text{v1} \\ \text{A2} & \text{v2} \end{bmatrix} \\ \text{ATTR2} & \text{---} \end{bmatrix} \quad \text{---} \quad \text{---} \\ \text{c.} & \begin{bmatrix} \text{ATTR1} & \begin{bmatrix} \text{A1} & \text{v1} \\ \text{A2} & \text{v2} \end{bmatrix} \\ \text{ATTR2} & f \end{bmatrix} \end{array}$$

Some grammatical phenomena, in particular coordination, adjunction and feature indeterminacy, are represented in LFG via set-valued attributes, as in (28).

$$(28) \quad f: \begin{bmatrix} \text{A} & \left\{ \begin{bmatrix} \text{DISTR1} & l \\ \text{DISTR2} & m \end{bmatrix}, \begin{bmatrix} \text{DISTR1} & l \\ \text{DISTR2} & n \end{bmatrix} \right\} \end{bmatrix}$$

Oleg Belyaev

At first sight, this may appear to violate the notion of f-structure as a function, and the consequent Uniqueness constraint (Section 3.4.1). However, sets in LFG are not multiple values of a single attribute; they are rather viewed as a special kind of f-structure – a *hybrid object* that has both attributes that pertain to it as a whole and attributes whose value is determined based on the values of the set members. This is based on the distinction between DISTRIBUTIVE and NON-DISTRIBUTIVE features.<sup>15</sup> The value of a *distributive* feature for a set is determined as follows:

- (29) If  $a$  is a *distributive* feature and  $s$  is a set of f-structures, then  $(s\ a) = v$  holds if and only if  $(f\ a) = v$  for all f-structures  $f$  that are members of the set  $s$ . (Bresnan et al. 1985; Dalrymple & Kaplan 2000)

Thus, a distributive feature for a set is only defined if it holds for all f-structures in the set. Thus, for (28), the equation  $(f\ A\ \text{DISTR1}) = L$  is true; conversely, no equation that invokes the feature DISTR2 (such as  $(f\ A\ \text{DISTR2}) = M$  or  $(f\ A\ \text{DISTR2}) = N$ ) can be satisfied, since the set elements differ in the value of this feature. Crucially, there is no requirement that distributive features be the same for all elements of a set unless they have been invoked; the structure in (28) is valid as long as the grammar does not assign any value to  $(f\ A\ \text{DISTR2})$ .

While distributive features are resolved on the basis of their values for individual members of a set, *non-distributive* features apply to sets as a whole:

- (30) If  $a$  is a *non-distributive* feature, then  $(f\ a) = v$  holds if and only if the pair  $\langle a, v \rangle \in f$ . (Bresnan et al. 1985; Dalrymple & Kaplan 2000)

In (3.3.3), the value of the attribute  $A$  illustrates a set with a non-distributive feature.

$$(31) \quad f: \left[ A \left[ \begin{array}{c} \text{NDISTR } N \\ \left\{ \begin{array}{c} \text{NDISTR } L \\ \text{NDISTR } M \end{array} \right\} \end{array} \right] \right]$$

This notation, standard in LFG work, is meant to represent that, while the feature NDISTR has the values  $L$  and  $M$  for the individual set members, it has the value  $N$  for the whole set. The equation  $(f\ A\ \text{NDISTR}) = N$  is therefore satisfied regardless of the set members' values of NDISTR.

<sup>15</sup>This distinction is normally understood as being grammar-wide, or even universal; some authors have recently proposed treating distributivity as a property of *feature application*, not features as such; the most recent such account seems to be Przepiórkowski & Patejuk (2012), and similar ideas are explored in Belyaev et al. (2015) and Andrews (2018).



## 2 Core concepts of LFG

Distributive and non-distributive features in LFG are used to model different ways in which feature values are resolved and checked in coordination and similar structures. For example, number is typically viewed as non-distributive, because a coordinate NP triggers plural agreement regardless of the number features of its conjuncts. In contrast, case is usually distributive: when a case value is assigned to a coordinate phrase, it must be borne by all its conjuncts. The issue of sets and distributivity with respect to coordination is dealt with in **chapters/Coordination**.

### 3.3.4 Semantic forms

A SEMANTIC FORM is a special type of attribute value that is exclusively assigned to the attribute PRED. Semantic forms consist of the predicate name followed by the list of its syntactic arguments; arguments that are assigned thematic roles are written in angled brackets, while arguments that are not thematic (such as expletive subjects or “raised” subjects and objects) are written outside angled brackets. For example, the PRED value for a transitive verb like ‘see’ will be ‘see<SUBJ OBJ>’. A verb like ‘rain’, which has no thematic arguments but an expletive subject, will have the PRED value ‘RAIN<SUBJ>’. Finally, an “object raising” verb like ‘believe’ will have the PRED value ‘BELIEVE<SUBJ>OBJ’: its subject is assigned a semantic role, while its object is not.

In the preceding paragraph, arguments were represented as mere lists of grammatical function names. This convention, which is followed in much LFG work (see e.g. Dalrymple 2001), is but a simplification: arguments inside PRED values are usually understood as direct references to the corresponding attribute values. Thus, in the left-hand side of (32), the PRED value is represented as ‘see<(*f* SUBJ) (*f* OBJ)>’. As observed in Kuhn (2003: 63), PRED values as used in typical LFG representations can be viewed as shorthands for complex structures such as in the right-hand side of (32);<sup>16</sup> FN is an abbreviation for FUNCTOR; SFID stands for “semantic form identifier”, on which see below. Similar structures are used in implemented parsers like the Xerox Grammar Writer’s Workbench (Kaplan & Maxwell 1996) and the Xerox Linguistic Environment (XLE, Crouch et al. 2008; see Forst & King 2021 [this volume]).

<sup>16</sup>I follow the representation used by Kuhn (2003), which does not distinguish between thematic and non-thematic arguments. In XLE, this is implemented by distinguishing between the attributes ARG1, ARG2, ... for thematic arguments and NOTARG1, NOTARG2, .... for non-thematic arguments.

Oleg Belyaev

$$(32) \quad \begin{array}{c} \left[ \begin{array}{c} \text{PRED} \text{ 'SEE'} \langle (f \text{ SUBJ}) (f \text{ OBJ}) \rangle \\ \text{SUBJ} \left[ \begin{array}{c} \text{PRED} \text{ 'JOHN'} \\ \text{NUM SG} \\ \text{PERS 3} \end{array} \right] \\ \text{OBJ} \left[ \begin{array}{c} \text{PRED} \text{ 'DAVID'} \\ \text{NUM SG} \\ \text{PERS 3} \end{array} \right] \end{array} \right] \\ f \end{array} = \begin{array}{c} \left[ \begin{array}{c} \text{PRED} \left[ \begin{array}{c} \text{FN SEE} \\ \text{ARGUMENT1} \\ \text{ARGUMENT2} \end{array} \right] \\ \text{SFID } i \\ \text{SUBJ} \left[ \begin{array}{c} \text{PRED} \left[ \begin{array}{c} \text{FN JOHN} \\ \text{SFID } j \end{array} \right] \\ \text{NUM SG} \\ \text{PERS 3} \end{array} \right] \\ \text{OBJ} \left[ \begin{array}{c} \text{PRED} \left[ \begin{array}{c} \text{FN DAVID} \\ \text{SFID } k \end{array} \right] \\ \text{NUM SG} \\ \text{PERS 3} \end{array} \right] \end{array} \right] \end{array}$$

If semantic forms were just a bundle of a functor and one or more argument slots, there would be no need to treat them as a special argument value type. What distinguishes them from any other value is their *uniqueness*: each introduction of a PRED value is treated as unique. That is, whenever an expression like  $(f \text{ PRED}) = \text{'FN'}$  introduces a new semantic form, it is assigned a unique identifier, even if it is lexically identical to another predicate. Thus the equivalence in (33): each PRED assignment is viewed as also introducing an invisible “index” to distinguish between individual PRED values. Thus, if atomic values can be introduced multiple times, PRED values cannot; different grammatical or discourse functions can have the same PRED value only through structure sharing,<sup>17</sup> when the whole f-structure is constrained to be identical. In XLE and other implemented versions of LFG, this uniqueness effect is achieved by including a special feature SFID in the PRED, that is assigned a unique value each time a PRED is introduced in the f-description.<sup>18</sup>

<sup>17</sup>Or, as an anonymous reviewer observes, by directly sharing the PRED value, although I am not aware of such analyses having been discussed.

<sup>18</sup>XLE extends standard LFG by allowing any atomic value to be unique – an *instantiated symbol* notated via a subscript following its name: VAL\_. Thus in XLE, semantic forms do not seem to require any special machinery as such. However, an anonymous reviewer observes that if the left-hand side of (32) is indeed the abbreviation of its right-hand side, it should be possible to manipulate argument structure in the syntax via equations such as  $(\uparrow \text{PRED ARGUMENT3}) = \downarrow$ . XLE seems to circumvent this by tacitly introducing a negative existential constraint that prevents any additional attributes from appearing in PRED except the ones included at its introduction. This includes both argument features and any other feature names: both the XLE version of the above statement and  $(\uparrow \text{PRED FOO}) = \text{BAR}$  lead to an existential constraint violation. It is also impossible to “construct” a semantic form using a set of separate statements for the individual features; thus even XLE does technically treat semantic forms as a special value type.

## 2 Core concepts of LFG

The uniqueness of PRED values is needed to prevent multiple introduction of arguments and will be discussed in Section 3.4.1.

$$(33) \quad \left\{ \begin{array}{l} (f \text{ PRED}) = \text{'APPLE'} \\ (f \text{ PRED}) = \text{'APPLE'} \end{array} \right\} \quad \equiv \quad \left\{ \begin{array}{l} (f \text{ PRED}) = \text{'APPLE}_1\text{' } \\ (f \text{ PRED}) = \text{'APPLE}_2\text{' } \end{array} \right\}$$

In current LFG research, PRED values mainly serve only to specify argument lists to satisfy Completeness and Coherence, and to provide unique “labels” for f-structures that have PREDs. Even this limited functionality is contested in the literature, with some authors proposing to abandon f-structures in favour of a purely semantic approach, see Section 3.4.4. Originally, however, PREDs were thought to have a more central role, providing a kind of link from syntax to semantics (Kaplan & Bresnan 1982). It is important to observe that PREDs are no longer viewed in these terms in the LFG literature; the functor names are only arbitrary labels, and all semantic derivation is separate from syntax, being done through Glue Semantics, described in **chapters/Glue**.<sup>19</sup>

### 3.4 Well-formedness conditions

There are three conditions that any f-structure must satisfy in order to be treated as valid: Uniqueness (also known as Consistency), Completeness, and Coherence. Any f-structure that violates these conditions cannot be part of a valid analysis of any sentence, regardless of the grammar as a whole.

#### 3.4.1 Uniqueness

##### 3.4.1.1 Definition

Uniqueness (Consistency) is the requirement that every attribute in an f-structure must have a single value. Thus, the two equations in (34) do not describe any valid f-structure.

(34) Ill-formed f-structure:

$$\begin{array}{l} (f \text{ A}) = \text{L} \\ (f \text{ A}) = \text{M} \end{array} \quad f\text{L} \left[ \begin{array}{c} \boxed{\text{L}} \\ \boxed{\text{M}} \end{array} \right]$$

<sup>19</sup> A kind of hybrid approach is proposed in Andrews (2008), which introduces a variant of Glue Semantics where meaning is at least in part derived from f-structure feature values; in this approach, PRED features do play a prominent role in semantic composition.

Oleg Belyaev

It should be noted that Uniqueness is not, in fact, a constraint that needs to be stipulated separately: it follows from the notion of f-structure as a function, since a function maps arguments to single values (thus defining a one-to-one or many-to-one, but not a one-to-many or many-to-many correspondence).

### 3.4.1.2 Multiple assignment

Uniqueness does not in any way imply that multiple assignments of an attribute value are ruled out. When the *same* value is assigned to an attribute two or more times, the resulting f-structure is valid, as seen in (35).

$$(35) \quad \begin{array}{l} (f \ A1 \ A2) = L \\ (f \ A1 \ A2) = L \\ (g \ A2) = L \\ (g \ A3) = M \\ (f \ A1) = g \end{array} \quad f \left[ \begin{array}{c} A1 \\ g \left[ \begin{array}{cc} A2 & L \\ A3 & M \end{array} \right] \end{array} \right]$$

In (35), the attribute ( $g \ A2$ ) is assigned its value three times and referred to in two different ways, but this “history” of its origin is not displayed in the resulting f-structure and is not recoverable from it in any way. This is an illustration of the LFG distinction between a *description* and the *object* that it describes, a crucial feature of LFG that separates it from most other frameworks, where syntactic constraints are usually encoded in the structure itself in one way or another.

Turning to a linguistically meaningful example, this distinction between description and object is manifest in the standard LFG approach to agreement (see **chapters/Agreement** for more detail). Agreement targets do not normally have a “copy” of their controller’s features; they only lexically specify the same features that are separately specified by the controller. If there is a conflict, the resulting f-structure is invalid. If there is no conflict, the agreement features are displayed in the f-structure once and there is nothing in the f-structure indicating that agreement feature checking has taken place. Compare the Italian examples (49) and (50) below, which map to the same f-structure even though the person-number features are described in two positions in (49) but defined once in (50).

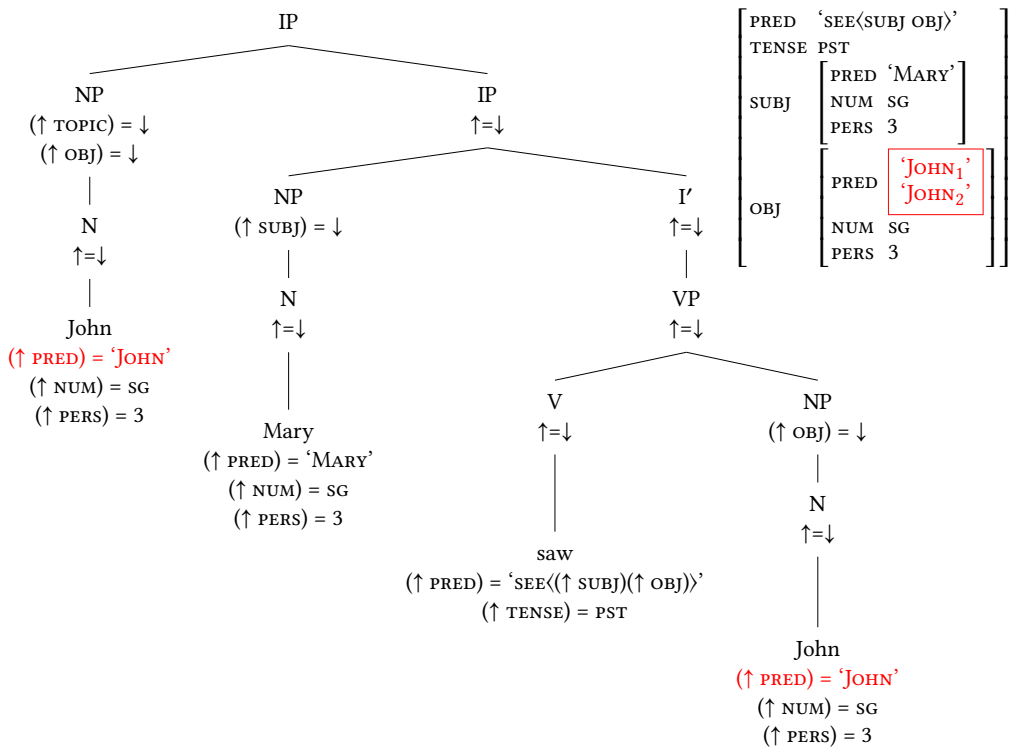
### 3.4.1.3 Uniqueness and PRED values

One place where multiple assignment is virtually prohibited is PRED features, whose values are special objects called semantic forms. As described in Section 3.3.4, each assignment of a PRED value is treated as a unique object; it is thus impossible to assign a PRED value more than once, even if the value to be assigned has the same functor name.

## 2 Core concepts of LFG

The reason why PRED values are treated in this way is to ensure that each argument position, and each predicative element in general, is instantiated by exactly one lexical head. Since there is no one-to-one correspondence between c-structure positions and f-structure functions, this cannot, in the general case, be ensured by phrase structure rules alone. Even in a configurational language like English, a displaced constituent is not directly linked to its “original” (normal, unmarked) position at c-structure; consequently, the c- to f-structure correspondence allows introducing it twice, as in (36).<sup>20</sup>

(36) Ill-formed f-structure for \**John, Mary saw John*:



What ensures the ungrammaticality of (36) is precisely the uniqueness of PRED values. This effect is even more pronounced in non-configurational languages, where no c-structure position is tied to any grammatical function, and any number of NPs may be freely mapped to any grammatical function; see **chapters/Cstr** for detail.

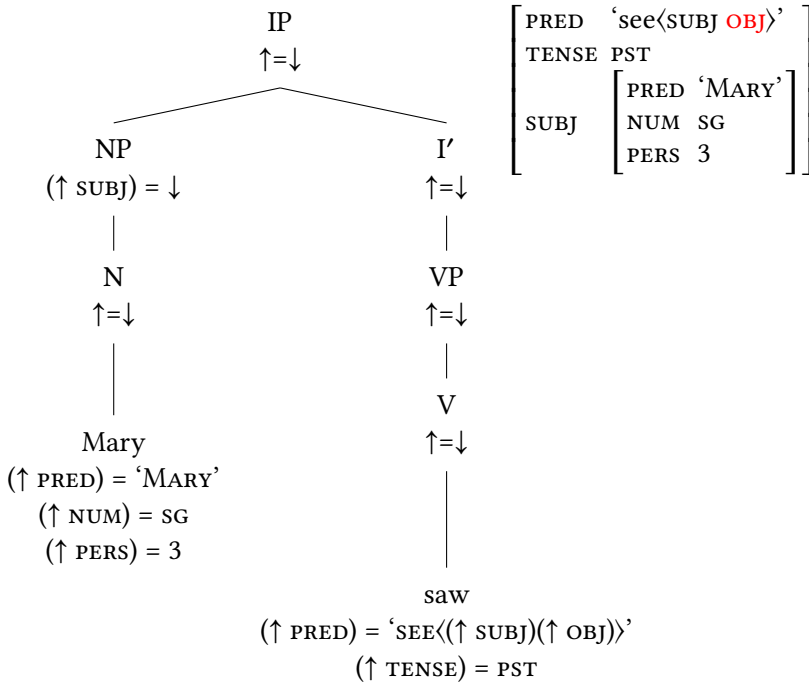
<sup>20</sup>For the sake of exposition, I assume that the topicalized direct object appears as an IP adjunct – this carries no theoretical significance.

Oleg Belyaev

### 3.4.2 Completeness

The Completeness condition requires every grammatical function governed by the PRED value of a given f-structure to exist in this f-structure. In other words, all arguments of a predicate must be “filled” by f-structures. This disallows examples such as (37).

(37) Ill-formed f-structure for *\*Mary saw*:



Recall that c-structure rules cannot be conditioned by argument structure; hence, Completeness violation is the only reason why this sentence is ungrammatical. The c- to f-structure correspondence is otherwise entirely valid.

It is important to understand that completeness only refers to f-structure and has nothing to do with whether arguments are expressed overtly or covertly. Since LFG avoids empty nodes, covert subjects in pro-drop languages do not correspond to any c-structure NP or DP, but Completeness still has to be satisfied at f-structure. This is normally done via equations introducing the pronominal PRED of the subject in the verb’s lexical entry: see (48) below.

An additional Completeness constraint has to do with the parameter of semantic argumenthood. It states that semantic arguments (i.e. those whose names

## 2 Core concepts of LFG

stand within angled brackets in the PRED) have to themselves contain a PRED. Conversely, non-arguments (those whose names stand outside angled brackets) are required not to contain a PRED, unless these f-structures are arguments or adjuncts elsewhere (such as, for example, in raising constructions). This is meant to exclude, respectively, expletive arguments in positions where semantic roles are assigned (38), and meaningful NPs in expletive positions (39).

(38) \*I saw **there**.

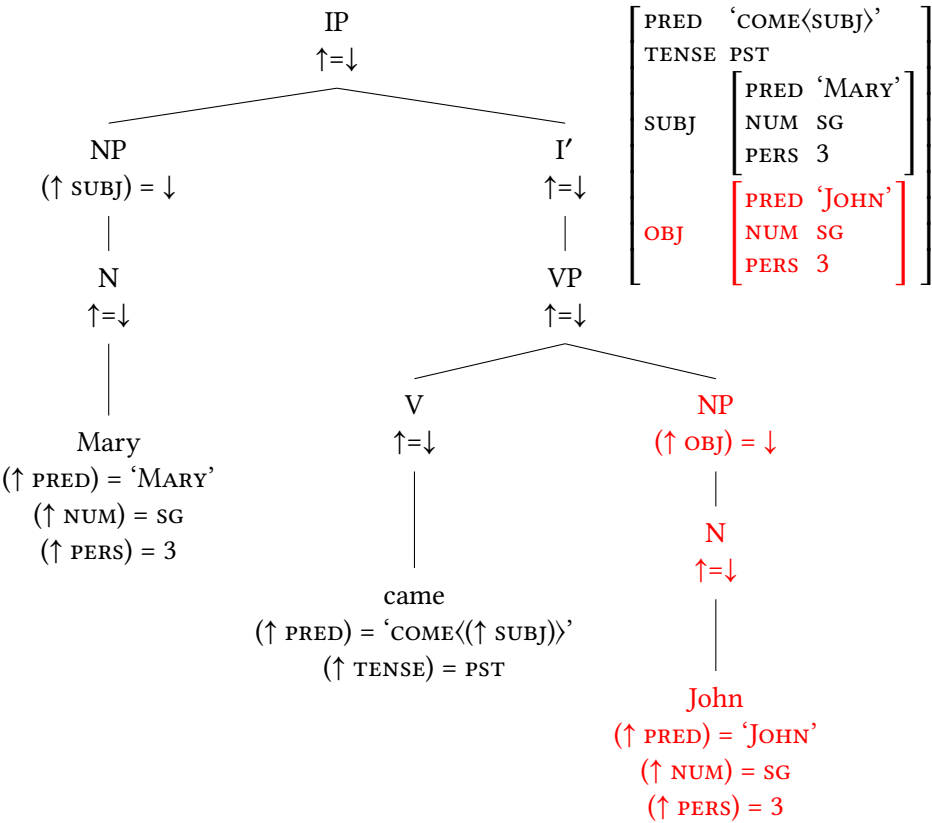
(39) \*The sky rained.

### 3.4.3 Coherence

The Coherence condition is the converse of Completeness: no governable functions (i.e. f-structure functions representing grammatical functions such as SUBJ, OBJ, etc., see **chapters/GFs**) may appear in an f-structure without being listed in a PRED value. This ensures that no “orphaned” arguments appear in an f-structure, disallowing examples such as (40).

(40) Ill-formed f-structure for \**Mary came John*:

*Oleg Belyaev*



Here, again, the c- to f-structure correspondence itself is valid, but the resulting f-structure is incoherent.

The coherence condition only applies to argumental grammatical functions and does not say anything about adjuncts or discourse functions. Where these elements may appear is constrained by a separate condition called Extended Coherence (Bresnan et al. 2016: 63). Extended Coherence requires that the f-structure where adjuncts appear have a PRED value. This ensures that no adjuncts appear in PRED-less f-structures. Discourse / overlay functions (DIS in more recent approaches, TOPIC and FOCUS in earlier work) are required to be linked to a grammatical function in some way: either functionally (via structure sharing) or anaphorically. For more information on the differences between various types of grammatical functions, see **chapters/GFs**.



### 3.4.4 Redundancy of PRED?

The description of Completeness and Coherence in this chapter follows the traditional LFG model, which had little to say about semantics; therefore, all valency restrictions had to be modeled at f-structure. Since at least the papers in Dalrymple et al. (1993), Glue Semantics (see **chapters/Glue**) has been gaining acceptance in LFG as the model of the syntax-semantics interface. Glue Semantics is resource-sensitive, which automatically ensures both Completeness and Coherence: Completeness, because all premises of the meaning constructor introducing the main predicate have to be saturated; Coherence, because no unused resources have to be left. The role of uniqueness of PRED for ensuring lack of multiple argument introduction / duplicate heads (Section 3.4.1) also follows from Glue semantics due to the fact that any resource can only be consumed once. Therefore, many authors, among others Kuhn (2001); Asudeh & Giorgolo (2012); Asudeh et al. (2014), have argued that PRED features in their original form are no longer necessary in LFG. At least argument lists can, for the most part, be safely dispensed with.<sup>21</sup> Many authors, therefore, continue to use PRED values but only include the name of the functor, not arguments in angled brackets; the remaining role of PRED values is only to provide an index for the f-structure, guaranteeing its uniqueness (that may be relevant for purely syntactic purposes that are not handled in semantics), and to provide information on the lexical content of its head.

## 4 The c- to f-structure mapping

### 4.1 Annotated c-structure rules

The metalanguage discussed in the preceding section can describe individual f-structures, but cannot, by itself, generate or evaluate natural language expressions. F-descriptions must come from somewhere. The only generative component in LFG is c-structure; therefore, phrase structure rules must be coupled with some mechanism that specifies how the nodes in the c-structure tree are mapped to f-structures – the projection function  $\phi$ . In LFG, this is normally done using ANNOTATED PHRASE STRUCTURE RULES where nodes at the right-hand side are supplemented by f-descriptions that reference the c- to f-structure mapping. This referencing is done by introducing two additional notational symbols:

---

<sup>21</sup>Non-thematic arguments like *it* in *it rained* might still be relevant insofar as they are not selected by any semantic predicate. However, these arguments may be forced to appear using existential constraining statements.

Oleg Belyaev

- (41) the current c-structure node:  $*$   
the immediately dominating c-structure node:  $\hat{*}$

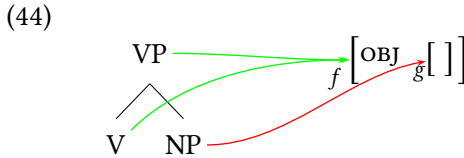
These are normally not used directly in LFG grammars; instead, two metavariables  $\downarrow$  and  $\uparrow$  are used, which signify the following:

- (42)  $\downarrow = \phi(*)$  (the f-structure corresponding to the current c-structure node)  
 $\uparrow = \phi(\hat{*})$  (the f-structure corresponding to the immediately dominating c-structure node)

This notation allows formulating rules of the type:

- (43)  $VP \rightarrow V \quad NP$   
 $\uparrow=\downarrow \quad (\uparrow OBJ)=\downarrow$

In (43), the annotation for V stands for “this node (V) maps to the same f-structure as the dominating node (VP)”, while the annotation for NP stands for “this node (NP) maps to the OBJ attribute of the f-structure of the dominating node (VP)”. The mapping that this rule defines is illustrated in (44). The nodes VP and V map to the same f-structure labeled as  $f$ , while NP maps to the f-structure labeled as  $g$  – the direct object of the clause.



The LFG metalanguage also allows for a notation for the inverse projection  $\phi^{-1}$ , that maps f-structures to the c-structure node(s) that map to them. This mapping is one-to-many and thus, unlike the direct projection, not a function. For example, in (44),  $\phi^{-1}(f)$  refers to two nodes: VP and V. The inverse projection is, by design, seldom used and, in fact, rarely required; but it is indispensable for certain construction which place selective requirements on the categorial status of their elements, such as the verb *wax* in examples like *wax poetical*, which is only compatible with an AP complement (the construction is discussed in Pollard & Sag 1994; for an LFG implementation, see Dalrymple et al. 2019: 6.10.3).

## 4.2 Some consequences of the mapping

### 4.2.1 Locality

The annotated rule format described in the preceding section is not merely a question of notation; it defines a rather rigid constraint on the way c-structure nodes can be mapped to f-structure. Namely, the mapping is strictly local: it can refer only to the nodes that are involved in a given phrase structure rule. It is not possible to freely traverse the tree and refer to, say, the node dominating the mother node, the child node of the current node, or the root node. LFG assumes that no linguistically meaningful generalizations can be captured using such “long-distance” references. For the majority of cases, this is clearly true, and consequently, there have been no serious attempts to extend the LFG meta-language in this direction.<sup>22</sup>

However, the strict locality of c- to f-structure mapping does create problems for the analysis of certain idiomatic combinations – multi-word expressions (MWEs), as they are called in the literature. Such MWEs often span whole syntactic phrases, and the lexical constraints involved cannot be captured locally. One solution that has been proposed in LFG is to replace the context-free c-structure by a variant of Tree-Adjoining Grammar (TAG), see Findlay (2017, 2019); this proposal is described in some detail in Section 5.2.

Within the local domain of c-structure rules, the mapping to f-structure is further constrained in that it is only possible to refer to the immediately dominating and current nodes, but not to any of the sister nodes. Unlike the locality constraint, this has been challenged in some LFG literature. For example, Dalrymple (2001: 120); Dalrymple et al. (2019: 222–223), developing the ideas of Nordlinger (1998), extends this notation by defining the metavariables  $<*$  and  $*>$  for “left sister of the current node” and “right sister of the current node”, respectively; the corresponding f-structures are  $\phi(<*)$  and  $\phi(*>)$ . Similarly, XLE defines the metavariables  $LS^*$  and  $RS^*$  for the same concepts.

The status of such innovations in the general LFG framework is uncertain. On the one hand, the analyses that introduce such notational conventions make convincing cases that they are necessary for analyzing certain phenomena, or at least vastly simplify such analyses. On the other hand, it is telling – and usually implied – that their use is somewhat exceptional and limited to a handful of specific phenomena. The fact that phrase structure nodes and lexical items do not

<sup>22</sup> As observed by an anonymous reviewer, if  $\uparrow$  and  $\downarrow$  are only abbreviations of  $\phi(\hat{*})$  and  $\phi(*)$ , it is possible to also use  $\phi(\hat{\hat{*}})$  and so on, making annotated rules potentially non-local. As mentioned above, low-level “designators” like  $*$  are not normally used in LFG analyses: grammars are expected to operate only with  $\uparrow$  and  $\downarrow$ .

*Oleg Belyaev*

refer to the information contributed by their left or right sisters in the vast majority of cases seems to be an important cross-linguistic generalization – one that is lost if this possibility is introduced in the formalism. If such formal devices are necessary, additional theoretical stipulations should supposedly constrain their use, but in practice, this possibility is almost never explored.

#### 4.2.2 Monotonicity

As Bresnan et al. (2016: 73ff.) observe, the limitations of the metalanguage described above (even if additional designations like  $\lt*$  and  $\gt*$  are included) leads to several important consequences for grammatical architecture. Specifically, the locality of the c- to f-structure mapping leads to the monotonicity of information flow in the syntax: the f-structure of a larger fragment is always more specific than the f-structure of a smaller fragment.

Let us first consider what “being more specific” means for an f-structure. By definition, f-structures are sets of feature-value pairs. It is clear, then, that  $g$  in (45) is more specific than  $f$ , as it has exactly the same features and values as  $f$  and one additional feature.

$$(45) \quad f \left[ \begin{array}{cc} A & X \\ B & Y \end{array} \right] \quad g \left[ \begin{array}{cc} A & X \\ B & Y \\ C & Z \end{array} \right]$$

Now consider a more complex case. In (46),  $f$  and  $g$  have the same features, but intuitively,  $g$  is more specific than  $f$  because the f-structure value of  $A$  in  $g$  is more specific than the value of  $A$  in  $f$ .

$$(46) \quad f \left[ \begin{array}{cc} A & [C \ Z] \\ B & Y \end{array} \right] \quad g \left[ \begin{array}{cc} A & \left[ \begin{array}{cc} C & Z \\ D & M \end{array} \right] \\ B & Y \end{array} \right]$$

Thus, specificity can be defined recursively:  $g$  is at least as specific as  $f$  if for every attribute  $a$  in  $f$ ,  $(ga) = (fa)$  or  $(ga)$  is at least as specific as  $(fa)$  (Bresnan et al. 2016: 74). This relation is essentially equivalent to subsumption (Dalrymple et al. 2019: 240), and can be notated accordingly:  $g \sqsupseteq f$  or  $f \sqsubseteq g$  means that  $g$  is at least as specific as  $f$ , or  $f$  subsumes  $g$ .

Now recall that every f-description can in principle correspond to infinitely many f-structures that satisfy it. Let us, then, define  $\phi(d)$  to be the smallest f-structure  $\phi$  that satisfies  $d$ ; this gives the mapping  $\phi$  from the set of functional descriptions  $D$  to the set of f-structures  $F$ . This mapping is MONOTONIC: the larger

## 2 Core concepts of LFG

the  $f$ -description  $d$ , the more specific the corresponding  $f$ -structure  $f$ . In other words,  $\phi : D \rightarrow F$  has the property that if  $d \subseteq d'$  and both  $d$  and  $d'$  have  $f$ -structure solutions, then  $\phi(d) \sqsubseteq \phi(d')$ .

This property of the mapping between  $f$ -descriptions and the corresponding  $f$ -structures follows from the nature of the  $f$ -structure equations: New equations can only specify additional information about the  $f$ -structure or check existing information; they cannot, as it were, “delete” existing feature values or otherwise make the structure less specific.

### 4.2.3 Fragmentability

Another feature of syntax in the LFG architecture that follows, in part, from monotonicity is **FRAGMENTABILITY** of language (Bresnan et al. 2016: 79–82). Recall that  $f$ -descriptions in annotated  $c$ -structure rules can only refer to the  $f$ -structures of the nodes involved in the rule (the node at the left side of the rule – the dominating node – and its daughters). This means that, the larger the tree, the longer its  $f$ -description; due to monotonicity, the  $f$ -structure of a larger tree fragment is, then, always more specific than the  $f$ -structure of any of its subtrees. Therefore, a valid  $f$ -structure can be constructed for any tree fragment dominating an arbitrary sequence of terminal nodes (a substring of a complete sentence), and this  $f$ -structure will not be overridden by any additional information that is contained in the complete sentence (unless it renders the  $f$ -structure ill-formed, in which case the sentence is ungrammatical).

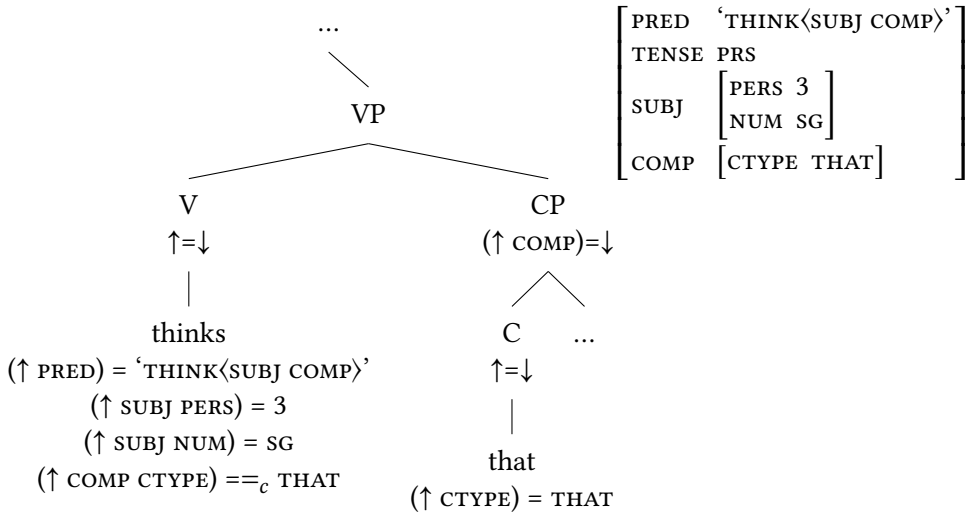
Note that this property of the  $c$ - to  $f$ -structure correspondence does not depend on whether the tree fragment corresponds to a sequence of terminal nodes in a complete sentence; it may even not be a constituent. Any sentence fragment is “self-contained” in the sense that its content is not modified by additional nodes in the tree.<sup>23</sup> Consider the  $c$ - and  $f$ -structures in (47). Here, the combination “thinks that”, which is not even a constituent in a fully formed sentence, contributes the argument structure of the matrix clause, the person and number features of the subject, and the complement type. It can be extended both upwards (with the addition of a subject) and downwards (with the addition of a complement clause), with  $f$ -structure information increasing monotonically in both cases.

---

<sup>23</sup>Note, however, that a tree fragment may be ambiguous between two or more interpretations; this ambiguity may be resolved by further material in the tree.

Oleg Belyaev

(47)



Within the non-transformational architecture of LFG, the properties of monotonicity and fragmentability may seem trivial. But this is not so for transformational frameworks, where elements may be extracted from within constituents, thus violating the principle of fragmentability: sentence fragments may become modified during derivation, losing some of the information they initially contain. Fragmentability captures the fact that sentence fragments frequently occur in natural discourse and are parsed without effort by native speakers.

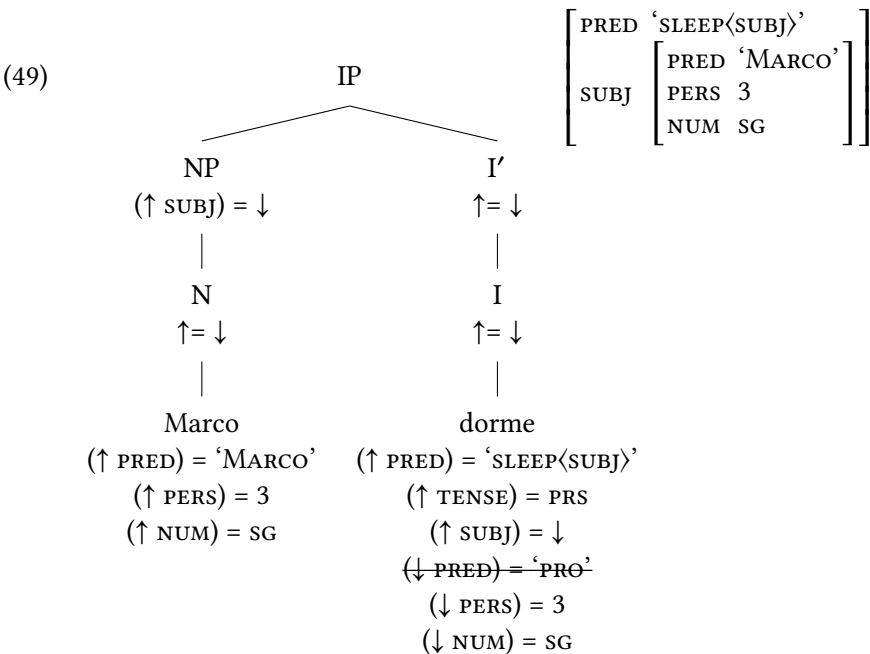
#### 4.2.4 Non-configurationality

Another consequence of the mapping between c- and f-structure is NON-CONFIGURATIONALITY of language. This property means that information in the f-structure does not necessarily correspond to specific positions in the tree. Thus, features of a single constituent may be “collected” from several nodes or assigned several times in different positions. This is usually related to the interaction between syntactic and morphological encoding.

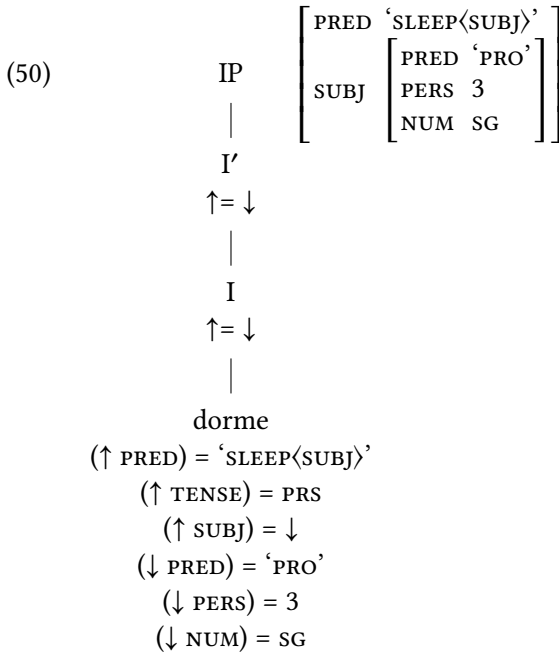
For example, in Italian, a pro-drop language, a third person singular verb form might be defined as in (48) – with the optional assignment of a PRED feature to the subject (the standard analysis of pro-drop in LFG). If there is a subject NP in Spec,IP, this annotation is not selected, as it would lead to a PRED conflict. If, however, there is no overt subject, this annotation *must* be used, because otherwise the resulting f-structure would violate Coherence: SUBJ would have no PRED value. Both options can be seen in (49) and (50).

## 2 Core concepts of LFG

- (48) *dorme*      V    (↑ PRED) = 'SLEEP<SUBJ>'  
                               (↑ TENSE) = PRS  
                               (↑ SUBJ) = ↓  
                                   ((↓ PRED) = 'PRO')  
                                   (↓ PERS) = 3  
                                   (↓ NUM) = SG



*Oleg Belyaev*

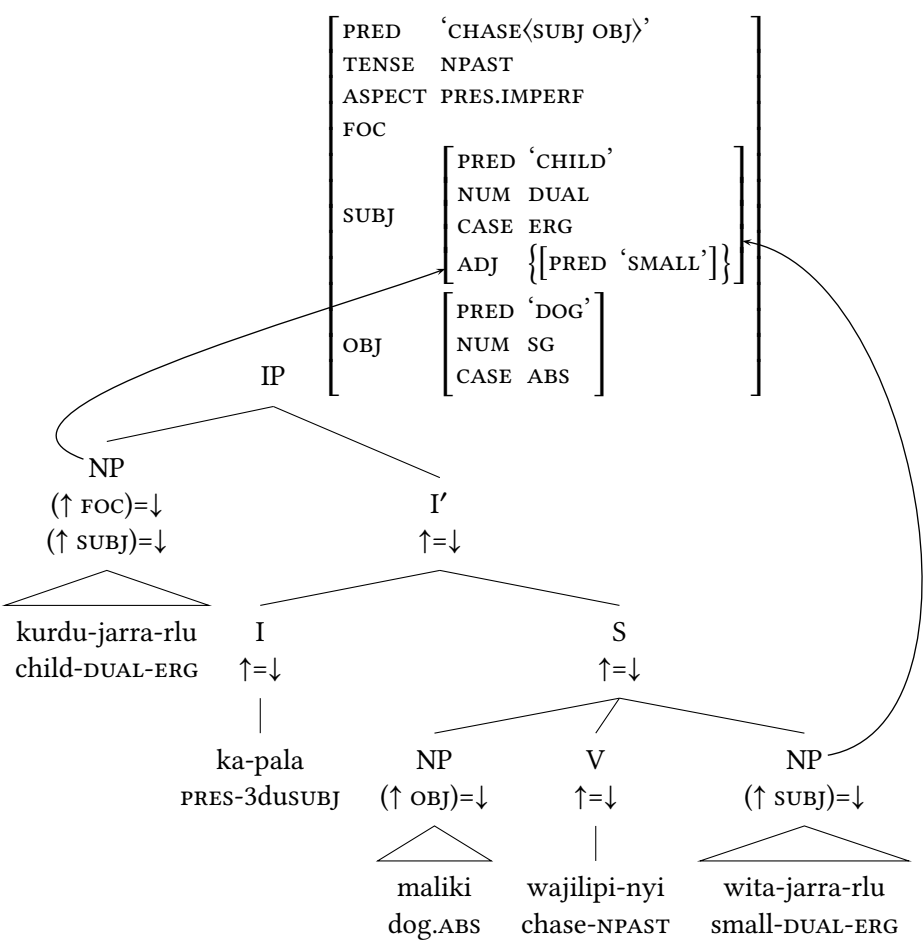


Thus, in Italian, the *PERS* and *NUM* features of the subject are always assigned at the *I* node, and they may also be assigned at the *N* head, if it is present. The *PRED* feature, in contrast, can be supplied either at the *I* node (if no overt head is present) or at the *N* node. This means that even in languages with relatively rigid word order and clausal phrase structure such as Italian (and English, although examples are less illustrative; see Bresnan et al. 2016), there is no universal mapping between c-structure positions and f-structure features.

“Non-configurationality” is usually understood in a more narrow sense, describing languages with no evidence for a hierarchical clause structure, such as Warlpiri (Hale 1983; Austin & Bresnan 1996). In (51), from Austin & Bresnan (1996: 229), two NPs, one having a head and the other only specifying an adjunct, map to the same f-structure function *SUBJ*. Thus information that is split at f-structure is collected together at f-structure.



(51)



These, of course, are only more radical manifestations of the phenomenon illustrated above. In Italian, the features of certain grammatical functions can be defined in different positions, but these positions, at least, are generally fixed, such that the overt subject, if present, occupies the Spec,IP position, the full NP direct object occupies Comp,VP, and the verb provides the agreement and PRED features of the subject. In radically non-configurational languages, in contrast, there is no association between c-structure positions and grammatical functions at all: any NP daughter of the S node can be mapped to any grammatical function, and any category, not only the verb, can function as the predicate of the clause. Non-configurational syntax and its challenges are described in more detail in **chapters/Cstr**.

*Oleg Belyaev*

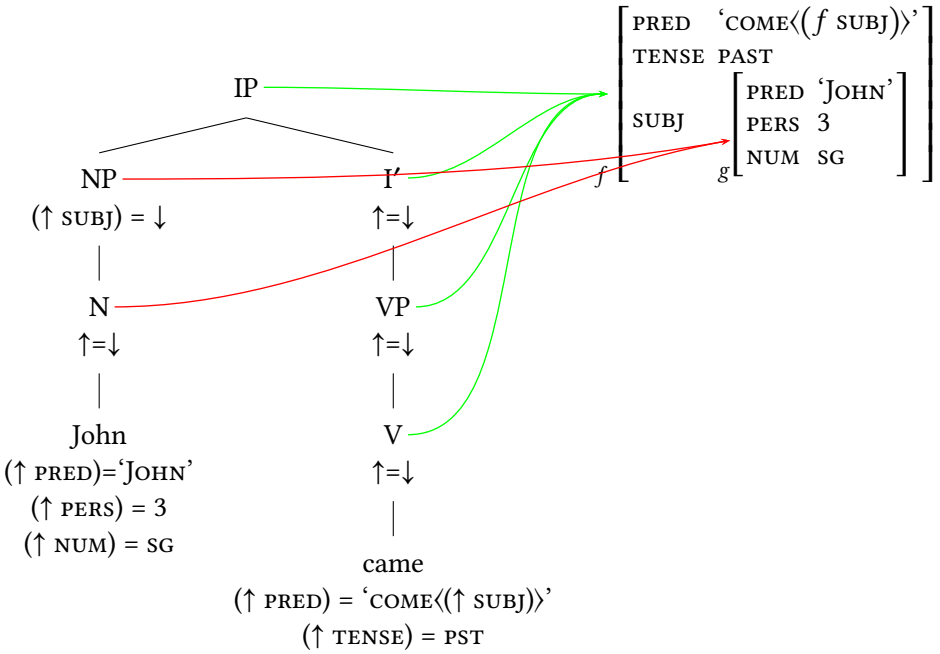
#### 4.2.5 Equality, unification, and non-compositionality

As seen in Belyaev 2021 [this volume] and elsewhere above, statements specifying the equality of one f-structure to another – most prominently,  $\uparrow=\downarrow$  – play a key role in the LFG c- to f-structure mapping and syntactic analyses. These kinds of statements allow mapping more than one c-structure node to the same f-structure and permit structure sharing and the checking of compatibility of f-structure features. Equality in LFG is very similar in its effects to UNIFICATION found in many other non-transformational formalisms – such that LFG itself is included in the class of UNIFICATION-BASED GRAMMARS in Shieber (1986).

However, as Kaplan (1989: 8ff.) points out, there is a crucial difference between LFG grammars and most unification-based frameworks (GPSG, HPSG, etc.): namely, the distinction between linguistic representations and the *descriptions* of said representations. The clearest case of this distinction are constraining equations, which impose additional constraints on admissible f-structures which, if not violated, do not show up anywhere in the f-structure. Defining equations behave similarly: the same feature may be defined several times in the tree, but the f-structure will contain no trace of its “pedigree”: only the resulting feature value will be included.

Another way in which LFG grammars are different from unification grammars is their NON-COMPOSITIONALITY. Even if a c-structure node is annotated with the “unificational” statement  $\uparrow=\downarrow$ , the f-structure it maps to in the complete sentence may contain additional values that are introduced higher in the tree. Thus, in (52) the VP node maps to an f-structure that includes a SUBJ feature that is not introduced anywhere in the VP subtree.

(52)



In a single-tier unificational model like GPSG or HPSG, where the counterpart to f-structure information directly occupies phrase structure nodes together with categorial information, the flow of information would be different: The content of a dominating node would be a function of the content of its children, hence, information contained in VP would be a subset of the information contained in IP. In LFG, as discussed above, *f-descriptions* do indeed increase monotonically, and a *fragment* associated with a node like VP does indeed contain a subset of the information contained in a larger constituent. However, in the full structure, this is not the case: every node mapped to a given f-structure maps to *all* the information contained in this f-structure, even to the information that is introduced only higher above.

### 4.3 Regularities in the c- to f-structure correspondence

In Section 2.1, I briefly described X' theory in the way that it is used in most LFG work. However, given that c-structure plays a limited role in LFG compared to the frameworks for which X' theory was originally devised, in this form it amounts to little more than a system for labelling nodes. In order to give significance to the notion of being a head, a specifier, a complement, or an adjunct, X'

*Oleg Belyaev*

Theory must be augmented by f-structure mapping principles.<sup>24</sup> A set of such principles is broadly accepted in LFG, although some details vary. For a more detailed exposition of  $X'$  theory, see Dalrymple et al. (2019); Bresnan et al. (2016).

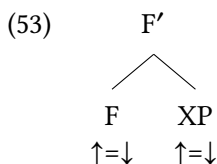
### 4.3.1 Heads

Headedness is a key concept of  $X'$  theory; preterminal nodes ( $X^0$ ) and all their projections ( $X'$  and  $X''$  in most versions of LFG) are heads. We saw in all examples above that all projections of a single  $X'$  category are mapped to the same f-structure, and this is for good reason:  $X$ -bar theory aims to model endocentricity, and so heads map to a “matrix” f-structure while specifiers, adjuncts, and complements (with the exception of functional categories) map to its dependents. Thus, heads are always annotated as  $\uparrow=\downarrow$ . This principle was first proposed in Bresnan (1982) and further developed in Zaenen (1983), where it is called the Head Convention.

This principle of head annotation allows us to formalize ENDOCENTRICITY as the requirement that every lexical category have a head (Bresnan et al. 2016), or, more correctly, an extended head (see below), because some phrases can have a lexically instantiated functional head but no lexically instantiated lexical head.

### 4.3.2 Complements

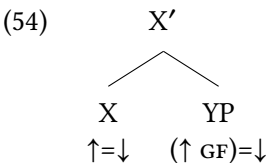
Complements are annotated differently depending on whether they are attached to functional or lexical heads. In essence, functional projections are little more than extensions of lexical projections, and generally map to the same f-structure: for example, CP, IP and VP map to the same clausal structure, while DP and NP map to the same nominal structure. Thus, complements of functional projections are f-structure CO-HEADS, annotated as  $\uparrow=\downarrow$  (53). The heads of functional categories are known as EXTENDED HEADS of lexical categories; a formal definition of extended head can be found in Bresnan et al. (2016: 136).



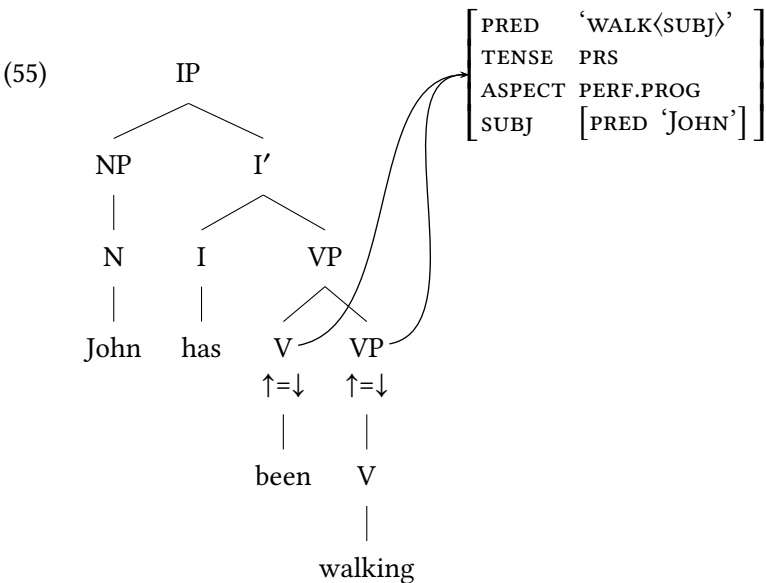
<sup>24</sup>It is by no means implied that these principles dictate the *only* annotations that can be associated with a given node: additional annotations are not only possible, but sometimes even required to produce a valid f-structure (for example, DIS must usually be associated with a grammatical function).

## 2 Core concepts of LFG

Complements of lexical projections are assigned to various functions of their heads' f-structures. Most typically these are, more specifically, grammatical functions, i.e. those functions that are governed by predicates and have no additional discourse significance (54); the label GF stands for "grammatical function" and includes such notions as subject (SUBJ), direct object (OBJ), secondary object (OBJ<sub>θ</sub>) and oblique (OBL<sub>θ</sub>). In Bresnan et al. (2016), this is formulated as a strict requirement that the complement may be any grammatical function except SUBJ (which, in their model, is both a grammatical and a discourse function, see **chapters/GFs**). However, this restricted understanding of lexical complements is not universally accepted. For example, Laczkó (2014) analyzes postverbal subjects in Hungarian as occupying the same position as postverbal direct objects, i.e. VP complements.



Complements of lexical heads may also behave in the same way as complements of functional projections, i.e. be annotated as  $\uparrow=\downarrow$ . This possibility should be allowed for to handle cases where the same f-structure extends over more than two projections, e.g. in certain English auxiliary constructions (55), see Bresnan et al. (2016).



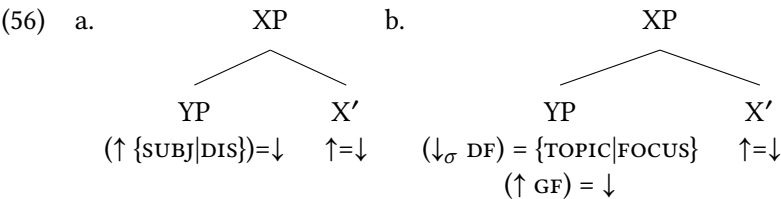
Oleg Belyaev

The higher VP in this case thus operates as a kind of intermediate functional projection. An alternative solution would be to introduce an additional functional projection for English, but this does not seem justified as the forms used in these positions are identical to V complements of simpler auxiliary constructions. At the same time, the X' model itself is obviously too simplistic to describe the full system of constraints on the English system of verbal periphrasis. This requires reference to morphological features of c-structure nodes; approaches that allow this will be briefly described in Section 5.

### 4.3.3 Specifiers

Specifiers are similar to complements in that they are mapped to f-structure positions in the f-structure of their heads. In the literature on LFG, there are two views on exactly what functions specifiers can be mapped to. The traditional approach as described in Dalrymple (2001); Bresnan et al. (2016) is that specifiers map to DISCOURSE FUNCTIONS (DF), which consist of TOPIC, FOCUS and SUBJ (which is unique in being simultaneously a grammatical function and a discourse function). However, a trend in much LFG work (King 1997; Butt & King 1997; Dalrymple & Nikolaeva 2011) is to eliminate information structure functions from syntax, instead relegating them to a separate projection, i-structure. Thus Dalrymple et al. (2019) instead propose that specifiers must be *either* syntactically prominent or prominent in information-structure terms. Syntactic prominence means that the f-structure of the specifier is either the subject, or it bears the overlay function DIS (which replaces the earlier TOPIC and FOCUS and handles long-distance dependencies). Discourse prominence means that the specifier occupies the discourse functions TOPIC or FOCUS at i-structure.<sup>25</sup> This question is discussed in more detail in **chapters/LDDs**.

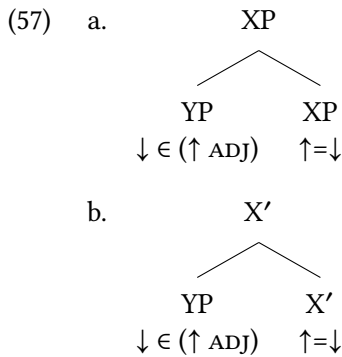
According to this approach, then, specifiers can be given annotations as in either (56a) or (56b):



<sup>25</sup>In contemporary LFG, discourse functions are usually modeled not in f-structure but in a separate projection, see Belyaev 2021: Section 5 [this volume] for the notation and **chapters/LDDs** for more information. The notation in (56) follows the model of information structure in Dalrymple & Nikolaeva (2011).

## 4.3.4 Adjunction

Unlike specifiers and complements, adjuncts may be freely iterated.<sup>26</sup> Naturally, then, they tend to be associated with the only grammatical function that is always set-valued,<sup>27</sup> ADJ (OR XADJ), see (57). As new adjuncts are added to the tree, they get added to the adjunct set, thereby not violating uniqueness.



C-structure adjuncts do not always map to f-structure adjuncts, however. Extraposed focused or topicalized material is often adjoined at c-structure, especially at XP level; it is then associated with an information structure function like TOPIC or FOCUS and with a grammatical function.

Some analyses also use adjunction as the main mechanism of introducing grammatical functions, not only adjuncts, into the f-structure, without them having any special information structure role. A prominent example is the analysis of Japanese and Korean in Sells (1994, 1995). Building on the ideas of Fukui (1986), Sells proposes that the maximal projection in Japanese and Korean is  $X'$ , and that the main sentence-building operation is the adjunction of verbal arguments and adjuncts to  $V'$ , and nominal dependents to  $N'$ . Adjunction of this sort can be described in LFG notation by rules such as (58), where GF is any grammatical function. Unlike flat structures of non-configurational languages, the resulting structures like (59), from Korean, are binary-branching, but the use of unrestricted adjunction of this kind ensures that the order of constituents is free.

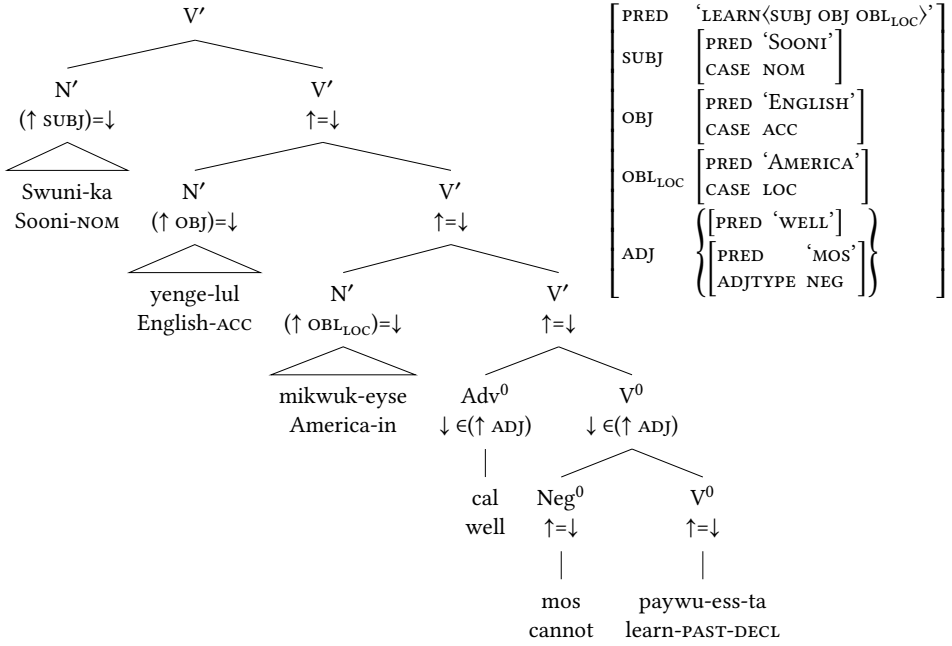
<sup>26</sup> As noted in Section 2.1, some versions of LFG  $X'$  theory allow multiple complements or specifiers. However, this is not the same as adjunct iterations, because, if multiple complements or specifiers are used in a grammar, these receive different annotations, thereby not causing a conflict. In contrast, multiple application of the same adjunct rule will lead to a uniqueness violation if it selects the same grammatical function.

<sup>27</sup> Due to the possibility of coordination, all grammatical functions can be set-valued. However, this requires the use of a special syntactic configuration at c-structure, whereas adjuncts are set-valued “by definition”.

Oleg Belyaev

$$(58) \quad X' \longrightarrow \begin{array}{c} Y' \\ (\uparrow \text{GF})=\downarrow \end{array} \quad \begin{array}{c} X' \\ \uparrow=\downarrow \end{array} \quad (\text{Sells 1994: 354})$$

(59)



'Sooni did not learn English well in America.' (Sells 1994: 355)

#### 4.3.5 The category S

As discussed above, the category S, being by definition exocentric, does not have a head in the  $X'$ -theoretic sense. This does not mean, however, that it has no head in the sense of c- to f-structure mapping, i.e. no node that is annotated as  $\uparrow=\downarrow$ . In fact, S usually includes at least one such node that represents the predicate; for example, in (7), representing the clause structure of Tagalog, the predicative XP is annotated as  $\uparrow=\downarrow$ , which causes the f-structure of the clause to be unified with the f-structure of the predicate, regardless of what its c-structure category may be.<sup>28</sup> Moreover, unlike  $X'$ -theoretic structures, a nonconfigurational S node can have more than one head: for example, a V node representing the lexical verb

<sup>28</sup>The actual developed analysis can be somewhat more complex, as there are several views on nonverbal predication in LFG, and the (non-)identity of its structure to that of verbal predications. For some discussion of this, see [chapters/Copula](#).



## 2 Core concepts of LFG

and an Aux node representing an auxiliary that contributes tense, agreement and other grammatical information.

It is remarkable that S is the only systematic exception from the X' schema<sup>29</sup> that is admitted in mainstream LFG, at least in theory. While the use of S for both nonconfigurational and “partially non-endocentric” languages like Tagalog or Irish is universally accepted as a valid and theoretically solid decision, there has been no discussion of exocentric NPs or other categories in the literature. Whether this represents a lack of empirical evidence for such structures in languages of the world, or is simply the result of a lack of focus and a kind of pre-determined conviction, is not clear.

### 4.3.6 Optionality of c-structure positions

Now that X' theory is supplemented by f-structure well-formedness constraints and annotation principles, we can introduce an additional feature of LFG c-structures: ECONOMY OF EXPRESSION, which amounts to optionality of most nodes, because the relevant grammatical constraints are for the most part captured at f-structure. This broad principle is formulated in the most radical way in Bresnan et al. (2016: 90), who state that *all* nodes (including nonbranching intermediate X' projection nodes, heads, complements and specifiers) are optional:

(60) ECONOMY OF EXPRESSION:

All syntactic phrase structure nodes are optional and are not used unless required by independent principles (completeness, coherence, semantic expressivity). (Bresnan et al. 2016: 90)

Note that this is a *theoretical* principle whose *formal* implementation is a separate issue, partly discussed in Section 5.2.1. For example, in the standard phrase structure rule formalism, the notions of complement and specifier crucially depend on the presence of intermediate X' nodes, even if these are redundant in the sense of unary branching. Thus, as Dalrymple et al. (2015) observe in their detailed discussion of economy of expression, this principle leads to a proliferation of rules, such as in (61).

(61) X' ELISION (Dalrymple et al. 2015: 384)

If an LFG grammar  $G_G$  contains an annotated rule of the form

---

<sup>29</sup>It is also the only consistent exception from endocentricity, although, as an anonymous reviewer observes, the X'-theory elaborated in Bresnan et al. (2016) only requires endocentricity for lexical, not functional, projections (p. 137), thereby allowing, among other things, the standard treatment of mixed categories (Bresnan et al. 2016: 311ff.).

*Oleg Belyaev*

$$\begin{array}{c}
 \text{XP} \longrightarrow \alpha \quad \text{X}' \quad \beta \\
 \qquad \qquad \qquad \uparrow=\downarrow \\
 \text{it also contains a rule of the form} \\
 \text{XP} \longrightarrow \alpha \quad \text{X} \quad \beta \\
 \qquad \qquad \qquad \uparrow=\downarrow
 \end{array}$$

In general, Dalrymple et al. (2015) conclude that economy of expression is plausible as an informal principle that emerges through the interaction of other, more basic principles, and that grammars, in general, tend to obey; but it is not plausible as a formal principle to be incorporated into the theory of grammar, because it not only introduces additional complexity into the framework, but also fails to account for cases of genuine non-optionality (such as, for example, in configurational languages where certain nodes are obligatory regardless of independent principles).

Still, the degree of optionality commonly allowed in LFG grammars is rather large and certainly greater than what is assumed by most other phrase-structure-based frameworks. I will now go through each of the X' theoretic categories and show why they can be optional (except adjuncts, because these are optional by definition, by virtue of the rules that introduce them).

#### 4.3.6.1 Complements and specifiers

Complements and specifiers not only can but must, as a rule, be optional because the c-structure does not contain any valency information and there is no way to verify at c-structure if, for example, the verb has a direct object. Thus, the rule in (43), repeated in (62), will hold for all English sentences, but the NP complement will only be licensed in transitive clauses.

$$\begin{array}{c}
 (62) \quad \text{VP} \longrightarrow \quad \text{V} \qquad \text{NP} \\
 \qquad \qquad \qquad \uparrow=\downarrow \quad (\uparrow \text{OBJ})=\downarrow
 \end{array}$$

If the verb is transitive (i.e. its PRED feature has OBJ in the list of arguments), omitting the complement will result in a violation of Completeness (unless the object is introduced in another position). By contrast, if the verb is intransitive, introducing the object here will lead to a Coherence violation, because the grammatical function OBJ will not be selected by any argument.

Optionality of complement and specifier positions, and c-structure positions where arguments are introduced in general, is also required because the material that they “canonically” contain may be displaced elsewhere, for example, to a position designated for wh-movement or information structure function. In this

## 2 Core concepts of LFG

case, only one position must be filled, otherwise conflict of PRED values will lead to a Uniqueness violation. Thus (43) may produce a single V node even in a transitive clause, provided that the direct object is introduced in another position (such as wh-movement *Whom did you see?* or topicalization *John, I saw*).

### 4.3.6.2 Heads

Similarly, c-structure heads can be optional in LFG because of Completeness and Coherence. PRED features are almost always<sup>30</sup> introduced by head nodes, i.e. nodes carrying the unificational annotation  $\uparrow=\downarrow$ . Therefore, a structure lacking a head (without its PRED features introduced elsewhere) will be PRED-less and will not be able to include any grammatical functions, because that would violate Coherence.

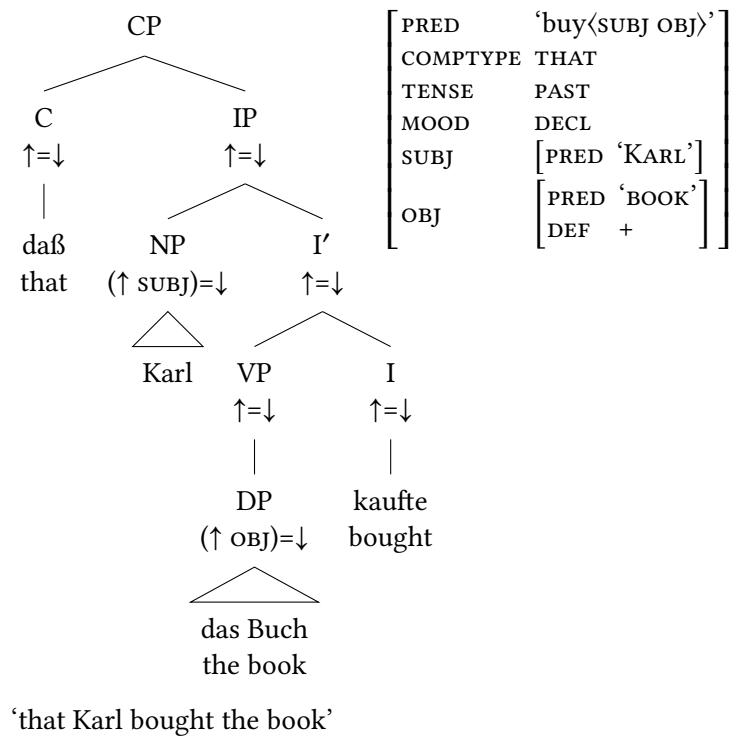
Headless XPs are quite widespread at clause level; their role is to account for variation in head positions in configurational languages. For example, in English lexical verbs always appear in V, but the I head can be filled or not depending on whether the verb form is periphrastic or synthetic. In German and other V2 languages, the distribution is more complex: the V head is only occupied if the verb form is periphrastic, and the auxiliary, or the finite verb in synthetic forms, stands in the I node in subordinate clauses (63) and in the C node in main clauses (64). Examples are from Bresnan et al. (2016: 448–450).

(63)

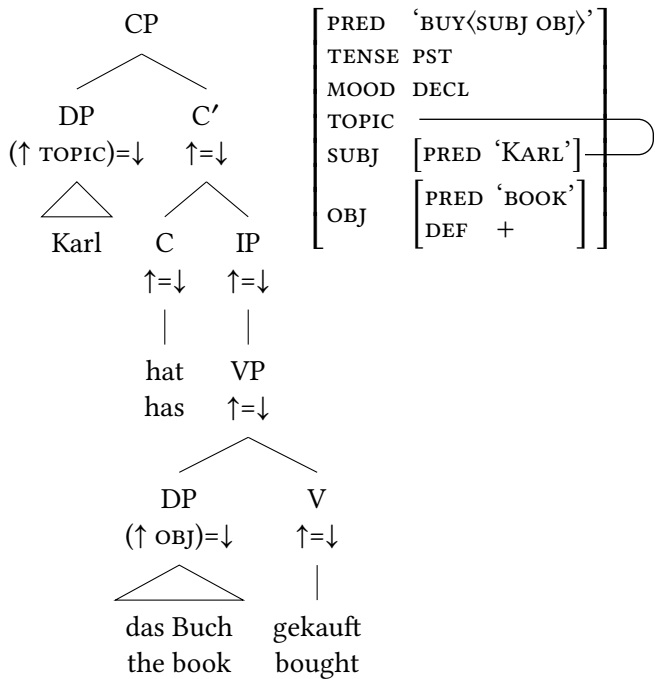
---

<sup>30</sup>It is technically possible to introduce a PRED feature in a different position. For example, the annotation of a complement or specifier might include an additional annotation like  $(\uparrow \text{OBJ PRED}) = \text{'PRO'}$ . I am not aware of any analyses utilizing this possibility; “external” PRED assignment normally only happens in verbal heads assigning PRED features to pro-dropped subjects and in similar such structures. However, Mary Dalrymple (p.c.) points out that such annotations seem to be required in asyndetic relative clauses like *The man John saw*, where the pronominal OBJ in the relative clause has to be introduced by a phrase structure rule since there is no lexical material that could plausibly contribute its content.

Oleg Belyaev



(64)



‘Karl has bought the book.’

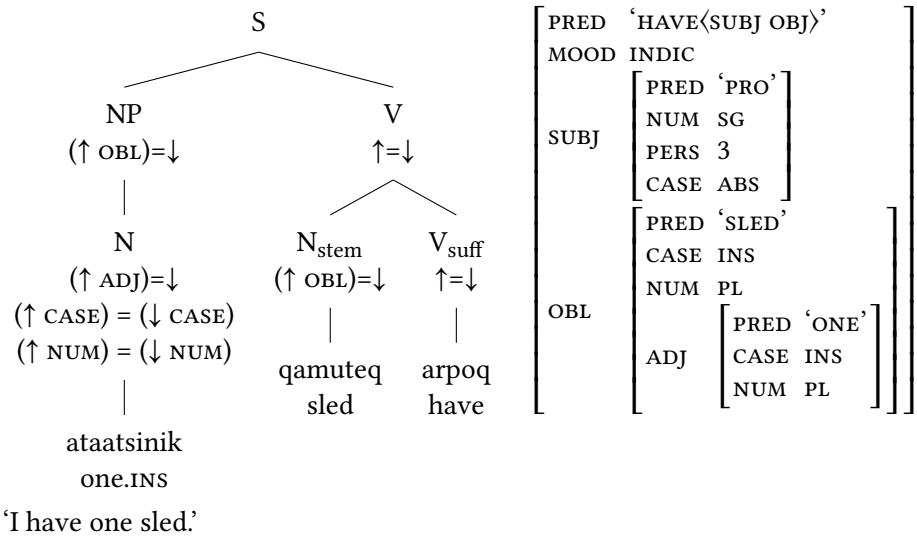
This analysis corresponds quite closely to the standard view of German word order in GB / Minimalism, such as Vikner (1995). The key difference is that there is no verb movement in LFG; verbs and auxiliaries are always “base-generated” in C, I, or V depending on clause types and the verb form. The correct word order is ensured by feature licensing; multiple occurrences of a verb form or verbless sentences are excluded at f-structure through Uniqueness and Coherence.

Another type of headless XP occurs in languages which allow freely discontinuous constituents, like the example from Warlpiri in (51) above. Non-configurational languages like Warlpiri allow freely assigning any grammatical function to Spec,IP (which is additionally interpreted as a focus) and to any NP children of S. Hence, two or more NPs might be mapped to the same grammatical function; if there is no PRED clash or case mismatch, the resulting sentences will be grammatical and these multiple NPs will be mapped to the same f-structure. For more information on non-configurational languages, see **chapters/Cstr**.

Finally, headless constituents appear in certain instances of incorporation, such as in West Greenlandic (65), where an incorporated noun head can have non-incorporated dependents (here, agreeing in instrumental case with the incorporated argument).

*Oleg Belyaev*

(65) Greenlandic (Bresnan et al. 2016: 446)



5 Extensions of the core architecture

The core architecture of LFG has remained remarkably stable since the framework was first introduced in Bresnan (1982); the only major innovations are the introduction of various additional projections, briefly described in Belyaev 2021: Section 5 [this volume], and functional uncertainty (earlier LFG used traces to model long-distance dependencies). Nevertheless, there have been proposals to alter and extend the core architecture, mainly from three directions: to adopt a view of c-structure different from context-free grammar; to introduce construction-based approaches to LFG using templates; to eliminate PRED values, fully relegating their work to semantics. None of these approaches have been adopted by mainstream LFG practitioners, with the exception of templates, which have gained some acceptance. Nevertheless, these proposals may represent venues in which LFG could develop in the future.

5.1 Constructions and LFG: Templates

In many ways, LFG is close in spirit to other non-transformational frameworks such as HPSG (Pollard & Sag 1994) or various versions of construction grammar (see Hoffmann & Trousdale 2013). All these frameworks, unlike mainstream generative grammar, are not committed to cross-linguistically universal structures

## 2 Core concepts of LFG

and instead define syntactic rules on a language-by-language basis. However, LFG is crucially different from these other approaches in lacking any concept comparable to the notion of construction. The basic building blocks of syntax are phrase structure rules and lexical entries (which formally are a subtype of phrase structure rules); there is a general set of principles governing the mapping from phrase structure positions to f-structure. It is, of course, possible to define separate phrase structure rules and lexical entries to handle specific phenomena and constructions, but these will not be formally related to other rules – there is no hierarchy of phrase structure rules that would allow defining, for example, an exceptional subtype of a specifier rule. In general, most theoretical principles in LFG (such as the principles of c- to f-structure mapping described above) are formulated in such a way as to define a structure that obtains by default, but which can be overridden in individual languages. This is at odds with the main tenets of construction-based approaches, where no general or universal principles or structures are usually assumed, and each construction hierarchy is language-specific.

Furthermore, while it is possible to define rules that are specific to individual constructions or lexical items, it is impossible to directly define a construction that spans more than the scope of one phrase structure rule (e.g., a specific combination of a specifier, head and complement). Of course, the same effect may be achieved by using combinations of defining and constraining equations, as, for example, in analyses of idioms; for an example, see Falk (2001: 77). But such analyses do not treat idioms or constructions as theoretical objects in their own right; the collocation is only enforced by the combination of equations acting at different levels.

These “limitations” related to the c-structure to f-structure correspondence are not necessarily disadvantages of the LFG system: they are the result of a conscious design decision that influences the way LFG analyses are structured; in most cases, it is possible to account for “construction-based” phenomena in LFG, but the description will be different than in Construction Grammar and related frameworks. However, there are certainly genuine cases of construction-specific phenomena, such as so-called multi-word expressions (MWEs); these are difficult to describe in standard LFG. A possible, but radical, solution is the replacement of context-free grammar by Tree-Adjoining Grammar (TAG) at c-structure, as described in [chapters/TAG](#).

Another reason why some counterpart to the notion of construction might be useful in LFG is that f-structure equations associated with rules and lexical items are not generalized in any way. Thus, nouns may have annotations such as  $(\uparrow \text{ NUM})=\text{SG}$  and  $(\uparrow \text{ NUM})=\text{PL}$ , and verbs,  $(\uparrow \text{ TENSE})=\text{PST}$ , but nothing in the

*Oleg Belyaev*

grammar *requires* nouns and verbs to introduce these equations, and there is no place where such generalizations are stated explicitly – in effect, they are only the result of consistency on the part of the grammar writer.<sup>31</sup> This limitation, again, cannot be overcome using a kind of type inheritance system common to construction-based approaches, because that would require a “hierarchy” of f-descriptions. But f-descriptions are only sets of expressions, not objects that can be manipulated or inherit information from each other.

A possible compromise between the description-based approach of LFG and constructions, explored in Asudeh et al. (2013), is based on the use of *TEMPLATES* – bundles of grammatical descriptions extensively used in computational LFG, such as in XLE, but also in some theoretical work (Dalrymple et al. 2004; Asudeh 2012). Templates are basically symbols that serve as shorthands for f-descriptions that are substituted for the template call wherever it is invoked in an f-description. For example, the combination of third person and singular number agreement, highly relevant for English grammar, can be abbreviated as the template 3SG (66). This template can then be called as in (67). Furthermore, just like an f-description, a template can be negated; thus, as Asudeh et al. (2013: 19) propose, English unmarked present-tense forms can be naturally captured as in (68a), which resolves to (68b).<sup>32</sup>

- (66) 3SG  $\equiv^{33}$   $(\uparrow \text{SUBJ PERS})=3$   
 $(\uparrow \text{SUBJ NUM})=\text{SG}$
- (67) *laughs* V  $(\uparrow \text{PRED}) = \text{'LAUGH<SUBJ>'}$   
 $@3\text{SG}$
- (68) a. *laugh* V  $(\uparrow \text{PRED}) = \text{'LAUGH<SUBJ>'}$   
 $\neg @3\text{SG}$
- b. *laugh* V  $(\uparrow \text{PRED}) = \text{'LAUGH<SUBJ>'}$   
 $\{ (\uparrow \text{SUBJ PERS}) \neq 3$   
 $| (\uparrow \text{SUBJ NUM}) \neq \text{SG} \}$

<sup>31</sup>Note that in LFG, this issue is distinct from the issue of permissible f-structure attributes and values discussed in Section 3.3. The two are, of course, related, and would have been the same issue in other frameworks, but not in LFG, where, as discussed above, structures are distinct from the descriptions that license them. An LFG grammar may not generalize over *structures* directly (unless feature declarations are used), but it may well generalize over *descriptions*.

<sup>32</sup>Note that such negation tacitly changes the equation type from defining to constraining, because negative statements can only be constraining. This change is not formally problematic, but care should be taken to ensure that other parts of grammar, which may depend on these defining equations, are not compromised.

<sup>33</sup>Asudeh et al. (2013) use  $:=$  for template assignment, which is a standard assignment operator in some programming languages (e.g. Pascal), also used in computer science.



## 2 Core concepts of LFG

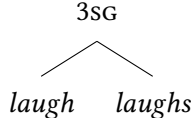
Templates can also be parametric, with parameters supplied in parentheses, as in programming languages. When a template is called, all mentions of each parameter are replaced by the string given in the parentheses. Note that this is done via simple string substitution,<sup>34</sup> and the parameters can be any kind of symbol; often, a reference to an f-structure, but not necessarily. For example, Asudeh et al. (2013) define the following template for intransitive verbs:

(69)  $\text{INTRANS}(P) \equiv (\uparrow \text{PRED}) = 'P\langle \text{SUBJ} \rangle'$

(70) *laughs*      V      @INTRANS(LAUGH)  
   @3SG

Templates by themselves are not theoretical objects: they are a simple mechanism for reusing common parts of f-descriptions. Nevertheless, if used consistently, they can serve as a powerful mechanism for capturing generalizations in grammatical structure. In particular, a kind of hierarchy of templates can be defined if the use of a template in a lexical item, phrase structure rule, or in another template is viewed as inheritance from that template. For example, both *laugh* and *laughs* inherit from the 3SG template:<sup>35</sup>

(71)



```

graph TD
    3SG[3SG] --- laugh[laugh]
    3SG --- laughs[laughs]
  
```

Asudeh et al. (2013) use this template system to develop a detailed analysis of the traversal / result construction (*Smithy drank his way through university*, Jackendoff 1992; Goldberg 1995) in English, Swedish, and Dutch. Since this seminal work, templates have been widely used in LFG literature, although their adoption is not universal. Importantly, an advantage of the template-based approach to constructions is that they only introduce a purely notational convention; they

<sup>34</sup>For this reason, if the parameter is an f-structure reference, it may be ambiguous within a template if it includes Functional Uncertainty. To ensure that the same f-structure is referred to in all expressions, the template should first assign the parameter to a local name.

<sup>35</sup>This may seem counterintuitive, given that *laugh* is not a third person singular form. However, inheritance in this approach is purely a matter of calling a template in the f-description: it does not matter in what context it is called (under negation, in disjunction, etc.). This graph captures the intuition that the English unmarked Present Simple form is defined with reference to the third person singular features (as opposed to, e.g., being a disjunction of all alternative person-number combinations). Note that “inheritance” here is purely a matter of visualization and metagrammatical analysis; it has no special status in the formalism itself.

*Oleg Belyaev*

do not change the architecture of LFG in any way. Thus template-based analyses are fully compatible with non-template-based ones.

This simplicity can also be perceived as a disadvantage, in that constructions are not “first class citizens” of the theory: the template mechanism is unconstrained, and its use is fully optional. However, this follows the overall spirit of LFG: As seen above, the core architecture and metalanguage are relatively unconstrained and certainly more expressive than is needed for the purposes of describing natural languages. Constraints on possible languages are meant to be captured by theoretical generalizations (such as the regularities of c- to f-structure mapping described in Section 4.3) that are not part of the formal framework itself. Likewise, templates only serve as a useful mechanism of generalizing over f-descriptions; what these templates should look like and how consistently they should be used are theoretical decisions that should be viewed as additional constraints on LFG grammars, not part of the formal architecture itself.

## 5.2 Modifications of c-structure

Compared to developments in other frameworks, such as Minimalism (cf. Adger 2013), there have been few advances in the development of constituent structure in LFG. Apart from the introduction of non-projecting words in Toivonen (2003), the version of X' theory used in most LFG work is the same as the original version developed in transformational grammar. However, there have been several alternative approaches to c-structure proposed in the literature, some relatively minor while others quite radical. In this section, I will describe two approaches – minimal c-structure (Lovestrand & Lowe 2017) and lexical sharing (Wescoat 2002). Another modification (Findlay 2017, 2019), which replaces context-free grammar with tree-adjoining grammar (TAG) while preserving core features of the LFG formalism, is described in **chapters/TAG**. Several categorial grammar-based approaches have been proposed (Oehrle 1999; Muskens 2001; Kokkonidis 2007), but have not gained much traction, possibly because they are no longer compatible with standard LFG and have to be regarded as separate, though related, frameworks.

### 5.2.1 Minimal c-structure

Lovestrand & Lowe (2017) propose a modification of X' theory to account for two shortcomings that they perceive in its standard LFG version. First, X' categories and projection levels are stipulated by the theory but not actually represented as discrete features; in formal terms, c-structure node labels are just monolithic

## 2 Core concepts of LFG

symbols, even though they are given a theoretical interpretation. Second, consistent application of  $X'$  theoretic principles leads to many redundant nodes, e.g. unary branching  $X'$  nodes have to be used if an  $XP$  has a complement but no specifier or adjuncts. This redundancy is sometimes eliminated by appealing to economy of expression, either by “pruning” the superfluous nodes (Bresnan et al. 2016) or by introducing additional rules into the grammar (such as  $XP \rightarrow X ZP$  in addition to  $XP \rightarrow X' YP$  and  $X' \rightarrow X ZP$ ). However, both solutions introduce additional complexity into LFG and could be avoided. Third, some analyses work with fewer than two levels of  $X'$  structure: for example, Bresnan et al. (2016: 130) take Welsh IP to lack a specifier, dominating only I and S. Sells (1994, 1995) similarly assumes that all phases in Japanese and Korean have  $X'$  as their maximal projection. This kind of “deficiency” is not formalized in traditional  $X'$  theory.

An earlier attempt to refine  $X'$  theory in LFG is Marcotte (2014), which, however, has been criticized in Lovestrand & Lowe (2017) for failing to account for some common syntactic structures, such as adjunction and non-projecting words. Lovestrand and Lowe propose, following Kaplan (1989), that additional categorial features are projected in a separate feature structure (l-structure) via the function  $\lambda$ . L-structure contains the features L (for Level) and P (for Projection) that represent the “current” bar level of the node and the maximal level that this particular phrase has in the sentence. C-structure itself only contains syntactic category information; thus  $X$ ,  $X'$ , and  $XP$  are all represented as  $X$ . Lovestrand and Lowe then define a set of templates and rule schemas that describe all the positions allowed by  $X'$  theory. For example, the template EXT in (72a) is a conjunction of the templates LPM (72b) and LP (72c), which mean that the annotated node is a maximal projection (LP) that is a daughter of a maximal projection (LPM). This applies to specifiers and adjuncts. The template HEADX (73a) is used on all  $X'$  theoretic heads and consists of the templates LDOWN (73b) and PUD (73c), which mean that, first, the bar-level of the annotated node is lower than the level of the mother by 1; (b) the maximal projection level is inherited from the head to the overall structure. These templates allow us to define the specifier rule template in (74).<sup>36</sup>

### (72) templates for specifier

- a. EXT  $\equiv$  @LPM  $\wedge$  @LP
- b. LPM  $\equiv$  ( $\hat{*}_\lambda$  L) = ( $\hat{*}_\lambda$  P)
- c. LP  $\equiv$  ( $^*_\lambda$  L) = ( $^*_\lambda$  P)

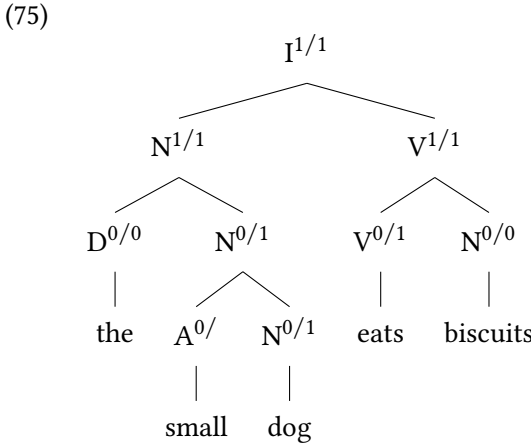
### (73) templates for head

<sup>36</sup>For clarity, conjunction is explicitly represented as  $\wedge$  in (72a) and (73a).

Oleg Belyaev

- a.  $\text{HEADX} \equiv @L\text{DOWN} \wedge @P\text{UD}$   
 b.  $\text{LDOWN} \equiv \{ (*_{\lambda} L) = 0 \wedge (\hat{*}_{\lambda} L) = 1 \mid (*_{\lambda} L) = 1 \wedge (\hat{*}_{\lambda} L) = 2 \}$   
 c.  $\text{PUD} \equiv (\hat{*}_{\lambda} P) = (*_{\lambda} P)$
- (74) specifier rule
- $$\begin{array}{ccc} X & \longrightarrow & Y \quad X \\ & & @EXT \quad @HEADX \end{array}$$

The application of this approach leads to c-structures notated as in (75), where the superscript numbers are shorthand for L/P feature values of the node.



In this example, prenominal A in English is treated as a non-projecting category, hence it lacks the *P* feature altogether.<sup>37</sup> It is seen from this example that the “maximal” projection level (*P*) is inherited bottom-up and represents the highest projection that the phrase has in this specific sentence. For example, the specifier noun phrase *the small dog* has a specifier, hence its dominating node has the category  $N^{1/1}$ , while the complement *biscuits* has no modifiers, and its head is only  $N^{0/0}$ . Thus the system results in minimal c-structures solely by using standard LFG mechanisms of templates and projections, without employing additional formal devices such as Economy of Expression.

### 5.2.2 Lexical sharing

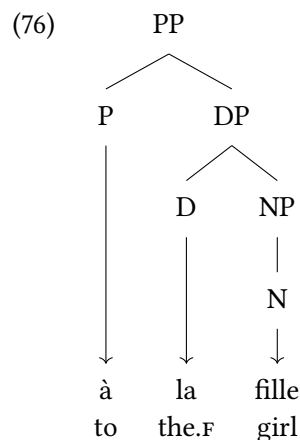
The principle of Lexical Integrity, and the general idea that there is a definite boundary between morphology and syntax, has long been criticized in the generative literature (perhaps the most recent such attempt is Bruening 2018) and,

<sup>37</sup>While Lovestrand and Lowe assume no DP in English, D is not treated as a non-projecting word: in their theory, *'s* possessors can attach to D as internal arguments (complements).

## 2 Core concepts of LFG

recently, in typological approaches (see Haspelmath 2011). Not all of the objections to lexicalism are necessarily applicable to LFG, but one persistent problem is the putative existence of syntactic structure where one lexical item (either completely idiosyncratic or derived in the morphology) occupies two or more syntactic heads. One example are preposition-determiner contractions in languages like French and German (Wescoat 2007): Items like French *au* [o] ‘to the (masculine)’ ( $\leftarrow \grave{a} + le$ ) are clearly idiosyncratic, historically motivated mergers of the preposition and the article (compare *à le faire* ‘to do it’, where *le*, identical in form to the masculine singular definite article, is the object proclitic of *faire* ‘do’, and thus does not trigger merger), but syntactically, they obey all the constraints that are independently imposed on prepositions and determiners in the language.

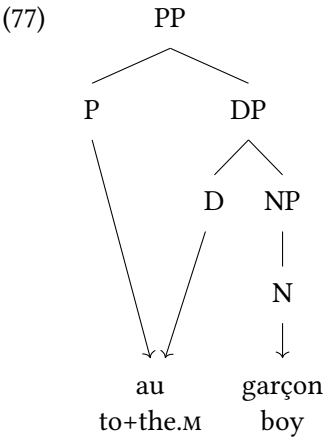
To account for such phenomena, Wescoat (2002) proposed **LEXICAL SHARING**: a modification of the LFG architecture to allow a single word (supplied by the lexicon) to occupy more than one c-structure node. In Wescoat’s system, lexical items are no longer part of c-structure; category nodes like N, V, I (preterminals in the standard system) are now terminal nodes that are mapped, via the projection function  $\lambda$ , to morphological words that comprise an ordered list at a separate level of representation, l-structure.<sup>38</sup> In the simplest and most common case, each terminal c-structure node corresponds to exactly one word:



Lexical sharing occurs when two or more terminal c-structure nodes are mapped to one morphological word:

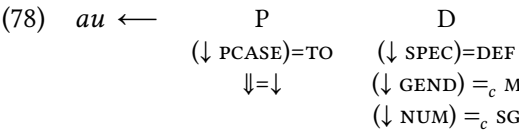
<sup>38</sup>It is unfortunate that the same name of the level and the projection function were independently used in Lovstrand & Lowe’s (2017) proposal of minimal c-structure, which creates confusion. However, as will be shown below, Wescoat’s approach can be integrated into the contemporary LFG architecture without stipulating an additional level.

Oleg Belyaev



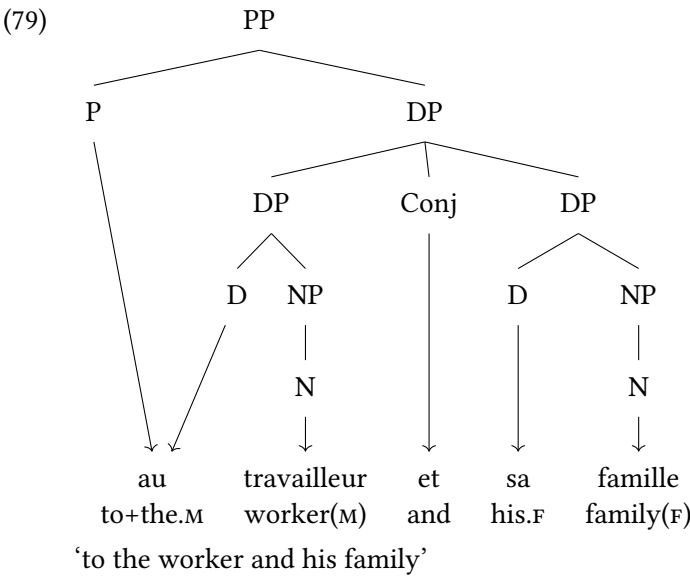
To avoid excessive reorderings, Wescoat puts a constraint on the correspondence between c-structure and l-structure which he calls the *order preservation axiom*: For all  $n_1$  and  $n_2$  in the set of terminal nodes, if  $\lambda(n_1)$  precedes  $\lambda(n_2)$ , then  $n_1$  precedes  $n_2$ . This means that words cannot be reordered. It also follows from this axiom that only adjacent nodes may be shared. Thus Lexical Sharing is, in fact, rather constrained and does not seem to introduce much additional complexity into the system.

Lexical entries in Lexical Sharing analyses are defined as in (78), with each node having a separate f-description. The syntactic analysis then proceeds according to the standard f-structure rules defined by the grammar; Lexical Sharing configurations are licensed if a word is defined as coinstantiating adjacent nodes.

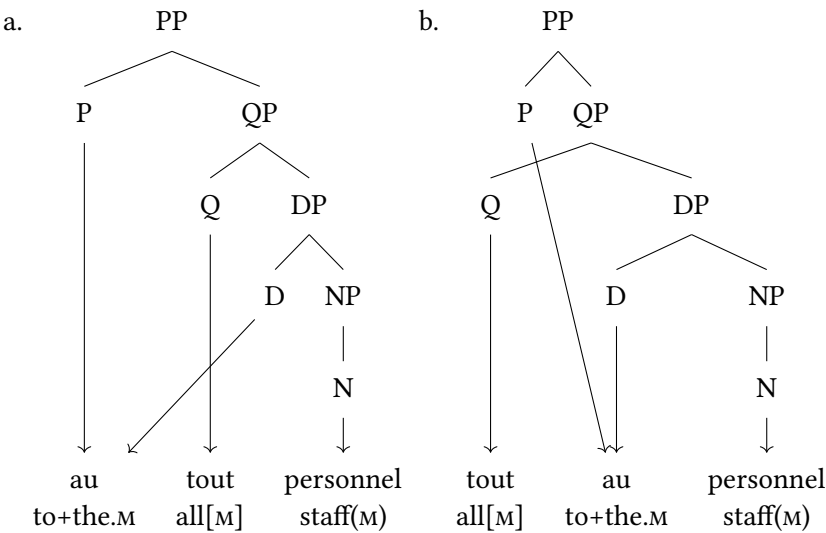


This correctly predicts the scope difference between the preposition and definite article in examples like (79). The order preservation axiom also predicts that structures like (80a) and (80b) are ungrammatical, because the shared nodes are not adjacent; the only possible word order is (81).

2 Core concepts of LFG



(80) Ill-formed c-structures:



(81) á tout le personnel  
to all[M] the.M staff  
‘to all the staff’

Note that Wescoat assumes that the correspondence function  $\phi$  should have l-structure in its domain, hence the use of  $\downarrow$  in annotations, instead of  $\uparrow$  in standard

*Oleg Belyaev*

LFG analyses. This assumption also motivates the symbol  $\Downarrow$ ; this stands for the abbreviation  $\phi(\lambda(\Downarrow))$ , i.e. “the f-structure of the lexical exponent of the current node” – this is needed to determine which of the co-instantiated f-structures the word itself maps to. However, this is not actually required, and Lowe (2016), in his analysis of the English “Saxon genitive” ’s, proposed a modification of lexical sharing that dispenses with both these additional notations and integrates the proposal into modern mainstream LFG. Lowe observes that Wescoat’s “l-structure” in fact serves the exact same function as the s-string – the set of morphosyntactic words that map to terminal c-structure nodes – in the LFG projection architecture, including the recent proposal of Dalrymple & Mycock (2011). Ordinarily, the s-string in LFG maps to terminal tree nodes that are occupied by morphosyntactic words; lexical sharing can be implemented by assuming that the c-structure tree terminates in category labels (preterminals), to which elements of the s-string are mapped. The replacement of l-structure by the s-string means that the symbol  $\Downarrow$  and all the related machinery is no longer needed, because s-structure does not map to f-structure.<sup>39</sup> For the same reason, lexical entries use  $\uparrow$ , as in normal LFG, instead of  $\downarrow$ .<sup>40</sup> In Lowe’s version of lexical sharing, the entry in (78) will look as follows:

- (82)    *au*: P D  
           P    ( $\uparrow$  PCASE)=TO  
           D    ( $\uparrow$  SPEC)=DEF  
               ( $\uparrow$  GEND) =<sub>c</sub> M  
               ( $\uparrow$  NUM) =<sub>c</sub> SG

While lexical sharing has been used to analyze several phenomena, including auxiliary reduction (Wescoat 2005), preposition-determiner contractions (Wescoat 2007), suspended affixation (Broadwell 2008; Belyaev 2014), endoclititics (Wescoat 2009), and morphologically bound complementation (Panova 2020), it has not been adopted as part of mainstream LFG, mainly, it seems, due to its apparent violation of Lexical Integrity and the potential to vastly increase the number of possible analyses. Indeed, if unconstrained, lexical sharing can be used to produce

<sup>39</sup>This seems rather harmless, because lexical sharing entries overwhelmingly just use  $\Downarrow=\downarrow$  on one of the nodes, which doesn’t seem to influence anything. However, Wescoat (2007) does use constraints on the l-structure to f-structure mapping to model certain limitations on preposition-determiner contractions in German.

<sup>40</sup>In fact, while standard LFG allows using  $\downarrow$  in lexical entries, this model does not. This means that analyses that make use of  $\downarrow$  in lexical entries, such as the Italian example in (48), have to be reformulated to use  $\uparrow$ . In most cases, this should not influence anything, although the definition and application of f-precedence might require some modification.



## 2 Core concepts of LFG

structures where every morphological category has its separate function projection that is shared with the lexical head, reminiscent of Distributed Morphology (DM, Halle & Marantz 1993). However, as both Broadwell (2008) and Lowe (2016) observe, lexical sharing can be constrained to be used only when there is independent syntactic evidence in favour of a separate lexical head. Under this interpretation, lexical sharing analyses present an advantage over DM analyses in that functional heads like CaseP or NumP are only stipulated as needed; for example, in Broadwell's (2008) analysis of suspended affixation, there is an empirical difference between languages where case is realized by a coinstantiated head (these allow suspended affixation) and languages where it is purely morphological (these do not); this opposition is lost in DM approaches, where other, arguably less intuitively plausible mechanisms have to be used, such as a constraint on coordinating sub-CaseP constituents, feature deletion (Kharytonava 2012), or morphological ellipsis (Erschler 2012).

Furthermore, as mentioned in Section 2.2, lexical sharing does not really violate Lexical Integrity as formulated in Bresnan & Mchombo (1995), see (8), i.e. as the general principle that words are built from different blocks and according to different rules than syntactic units. Indeed, syntax does not have any access to internal word structure in lexical sharing analyses, and coinstantiated nodes map to words as complete units, not to morphemes, disembodied features, or anything similar.<sup>41</sup> This gives lexical sharing analyses a distinct flavour that separates them from both mainstream LFG analyses and from truly non-lexicalist analyses that situate morphemes or features in functional projections; notably, it still allows an independent morphological module (usually described in LFG in terms of a lexicalist realizational framework like PFM, see **chapters/Morphology**) to do its work.

## 6 Conclusion

In this article, I have tried to summarize the state of the art of the core syntactic representations of LFG – the c- and f-structures, while also providing information on how the understanding of various phenomena evolved over time. From

---

<sup>41</sup>In fact, from a certain perspective this might be viewed as a disadvantage of lexical sharing analyses in that they fail to capture the fact that coinstantiated material usually corresponds to a well-defined, segmental, agglutinatively attached element of the wordform. For example, Ossetic affixal case features are realized on the Case head, while stem-based ones are realized on N (Belyaev 2014). I am not aware of any analyses where coinstantiated heads encode features that are realized by stem change, suppletion, or apophony. This fact might be explained diachronically, however, since lexical sharing usually reflects an ongoing process of (de)grammaticalization.

*Oleg Belyaev*

this exposition, it can be seen that while this understanding has considerably changed in almost all areas of grammar (for example, in the understanding of subjecthood and overlay functions), the formal underpinnings of LFG have remained remarkably stable over the years. The only fundamental innovation to the original c- and f-structure architecture of Kaplan & Bresnan (1982) is the introduction of functional uncertainty in Kaplan & Zaenen (1989b). Since then, new levels of projection were introduced, and the architecture extended in various ways, but the core mechanisms of c- and f-structure – notation, featurehood, even the basic set of GFs – have remained constant. This serves as an impressive testimony of the versatility of the architecture proposed in Kaplan & Bresnan (1982), and its remarkable suitability to describing natural languages.

The architecture of LFG is both similar to that of other constraint-based frameworks and very different from them in various ways. The main difference is the parallel architecture of LFG, and the related emphasis on the distinction between descriptions (a set of syntactic constraints) and the structures that are licensed by these descriptions. While constructions and lexical entries are *structures* in most other frameworks, in LFG they are sets of *statements* that describe a range of possible structures. This architectural feature enables LFG to make use of mechanisms such as functional uncertainty and inside-out application, which are unavailable in other frameworks. LFG is also special in that it can be viewed as incorporating the best features of constituent-structure-based (at c-structure) and dependency-based (at f-structure) frameworks, while avoiding their main drawbacks. Frameworks that use phrase structure as the only syntactic representation require additional mechanisms such as transformations, multiple dominance or separate linearization to properly capture word order variation and feature constraints; LFG manages to keep c-structure relatively simple due to the fact that all feature interactions are captured at f-structure, without referring to constituent structure positions. At the same time, the fact that f-structure does not directly refer to individual words or phrase structure nodes allows adequately capturing word order variation while keeping predicate-argument representations fairly uniform across languages.

While the empirical coverage of LFG work is impressive, and a number of important developments are now taking place in several theoretical directions, not all areas of syntax have been researched to the same extent. The focus on f-structure and the view of GFs as theoretical primitives has prompted a lot of fruitful and insightful work on subjects and other core grammatical relations. Functional uncertainty and structure sharing have also proved to be efficient mechanisms for describing long-distance dependencies. The notion of sets and

## 2 Core concepts of LFG

feature distributivity allow for elegant analyses of coordination – an area traditionally underrepresented in mainstream syntactic frameworks. In contrast, c-structure has seen much less attention,<sup>42</sup> although here important developments are also taking place. The notion of lexical integrity, assumed as a stipulation early in the history of LFG, has not been extensively discussed and refined, in spite of numerous challenges. These challenges will have to be dealt with if LFG is to compete with other frameworks for the originally envisaged role of “a theoretically justified representation of the native speaker’s linguistic knowledge” (Kaplan & Bresnan 1982).

## Acknowledgements

This research has been supported by the Interdisciplinary Scientific and Educational School of Moscow University “Preservation of the World Cultural and Historical Heritage”.

## References

- Adger, David. 2013. *A syntax of substance* (Linguistic Inquiry Monographs 64). Cambridge, MA: The MIT Press. DOI: 10.7551/mitpress/9780262018616.001.0001.
- Andrews, Avery D. 2008. The role of PRED in LFG+Glue. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 46–67. Stanford, CA: CSLI Publications.
- Andrews, Avery D. 2018. Sets, heads and spreading in LFG. *Journal of Language Modelling* 6(1). 131–174. DOI: 10.15398/jlm.v6i1.175.
- Asudeh, Ash. 2012. *The logic of pronominal resumption* (Oxford Studies in Theoretical Linguistics). Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199206421.001.0001.
- Asudeh, Ash, Mary Dalrymple & Ida Toivonen. 2013. Constructions with lexical integrity. *Journal of Language Modelling* 1(1). 1–54. DOI: 10.15398/jlm.v1i1.56.

---

<sup>42</sup>The reason for this might be that the range of phenomena handled by c-structure is much less than those handled by f-structure, as c-structure only directly models word order and embedding. However, as an anonymous reviewer observes, c-structure in LFG is analogous to Merge in Minimalism, being the main generative component that connects different projections together while also providing codescription for the semantics. This role can hardly be described as minor, but the existing model handles this purpose rather adequately.

*Oleg Belyaev*

- Asudeh, Ash & Gianluca Giorgolo. 2012. Flexible composition for optional and derived arguments. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 64–84. Stanford, CA: CSLI Publications.
- Asudeh, Ash, Gianluca Giorgolo & Ida Toivonen. 2014. Meaning and valency. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 68–88. Stanford, CA: CSLI Publications.
- Asudeh, Ash & Ida Toivonen. 2006. Symptomatic imperfections. *Journal of Linguistics* 42. 395–422.
- Austin, Peter K. & Joan Bresnan. 1996. Non-configurationality in Australian aboriginal languages. *Natural Language & Linguistic Theory* 14. 215–268. DOI: 10.1007/bf00133684.
- Belyaev, Oleg. 2013. Optimal agreement at m-structure. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '13 conference*, 90–110. Stanford, CA: CSLI Publications.
- Belyaev, Oleg. 2014. Osetinskij kak jazyk s dvuxpadežnoj sistemoj: gruppovaja fleksija i drugie paradoksy padežnogo markirovanija [Ossetic as a two-case language: suspended affixation and other case marking paradoxes]. *Voprosy jazykoznanija* 6. 31–65.
- Belyaev, Oleg. 2021. Introduction to LFG. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 3–20. Berlin: Language Science Press. DOI: ??.
- Belyaev, Oleg, Mary Dalrymple & John J. Lowe. 2015. Number mismatches in coordination. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '15 conference*, 26–46. Stanford, CA: CSLI Publications.
- Blackburn, Patrick & Claire Gardent. 1995. A specification language for Lexical-Functional Grammars. In *Proceedings of the 7th conference of the European chapter of the ACL (EACL 1995)*, 39–44. European Association for Computational Linguistics. DOI: 10.3115/976973.976980.
- Bögel, Tina, Miriam Butt, Ronald M. Kaplan, Tracy Holloway King & John T. Maxwell III. 2010. Second position and the prosody-syntax interface. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 106–126. Stanford, CA: CSLI Publications.
- Bresnan, Joan. 1982. Control and complementation. *Linguistic Inquiry* 13(3). 343–434.
- 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, Ronald M. Kaplan & Peter Peterson. 1985. Coordination and the flow of information through phrase structure. Unpublished manuscript, Xerox PARC.

## 2 Core concepts of LFG

- Bresnan, Joan & Sam A. Mchombo. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language & Linguistic Theory* 13(2). 181–254. DOI: 10.1007/bf00992782.
- Broadwell, George Aaron. 2008. Turkish suspended affixation is lexical sharing. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 198–213. Stanford, CA: CSLI Publications.
- Bruening, Benjamin. 2014. Precede-and-command revisited. *Language* 90(2). 342–388. DOI: 10.1353/lan.2014.0037.
- Bruening, Benjamin. 2018. The lexicalist hypothesis: Both wrong and superfluous. *Language* 94(1). 1–42. DOI: 10.1353/lan.2018.0000.
- Butt, Miriam & Tracy Holloway King. 1997. Null elements in discourse structure. Unpublished manuscript intended for publication in the Proceedings of the NULLS Seminar. <http://ling.uni-konstanz.de/pages/home/butt/main/papers/nulls97.pdf>.
- Carpenter, Robert L. 1992. *The logic of typed feature structures*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/cbo9780511530098.
- Chomsky, Noam. 1970. Remarks on nominalization. In Roderick A. Jacobs & Peter S. Rosenbaum (eds.), *Readings in English transformational grammar*, 184–221. Waltham, MA: Ginn.
- Corbett, Greville G. 2012. *Features*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/cbo9781139206983.
- Crouch, Richard, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III & Paula Newman. 2008. *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).
- Crouch, Richard & Tracy Holloway King. 2008. Type-checking in formally non-typed systems. In *Proceedings of the ACL workshop on software engineering, testing, and quality assurance for natural language processing*, 3–4. DOI: 10.3115/1622110.1622112.
- Dalrymple, Mary. 1993. *The syntax of anaphoric binding*. Stanford, CA: CSLI Publications.
- Dalrymple, Mary. 2001. *Lexical Functional Grammar*. Vol. 34 (Syntax and Semantics). New York: Academic Press. DOI: 10.1163/9781849500104.
- Dalrymple, Mary & Ronald M. Kaplan. 2000. Feature indeterminacy and feature resolution. *Language* 76. 759–798. DOI: 10.2307/417199.
- Dalrymple, Mary, Ronald M. Kaplan & Tracy Holloway King. 2004. Linguistic generalizations over descriptions. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*, 199–208. Stanford, CA: CSLI Publications.

Oleg Belyaev

- Dalrymple, Mary, Ronald M. Kaplan & Tracy Holloway King. 2015. Economy of Expression as a principle of syntax. *Journal of Language Modelling* 3(2). 377–412. DOI: 10.15398/jlm.v3i2.82.
- Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III & Annie Zaenen (eds.). 1995. *Formal issues in Lexical-Functional Grammar*. Stanford, CA: CSLI Publications.
- Dalrymple, Mary, John Lamping & Vijay Saraswat. 1993. LFG semantics via constraints. In *Proceedings of the 6th conference of the European chapter of the ACL (EACL 1993)*, 97–105. DOI: 10.3115/976744.976757.
- Dalrymple, Mary, John J. Lowe & Louise Mycock. 2019. *The Oxford reference guide to Lexical Functional Grammar*. Oxford: Oxford University Press. DOI: 10.1093/oso/9780198733300.001.0001.
- Dalrymple, Mary & Louise Mycock. 2011. The prosody-semantics interface. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '11 conference*, 173–193. Stanford, CA: CSLI Publications.
- Dalrymple, Mary & Irina Nikolaeva. 2011. *Objects and information structure* (Cambridge Studies in Linguistics). Cambridge, UK: Cambridge University Press.
- Erschler, David. 2012. Suspended affixation in Ossetic and the structure of the syntax-morphology interface. *Acta Linguistica Hungarica*. DOI: 10.1556/aling.59.2012.1-2.7.
- Everett, Dan. 2015. Review of *A Syntax of Substance* by David Adger. *American Anthropologist* 117(2). 414–449. DOI: 10.1111/aman.12251.
- Falk, Yehuda N. 2001. *Lexical-Functional Grammar: An introduction to parallel constraint-based syntax*. Stanford, CA: CSLI Publications.
- Falk, Yehuda N. 2010. An unmediated analysis of relative clauses. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 207–227. Stanford, CA: CSLI Publications.
- Findlay, Jamie Y. 2017. Multiword expressions and lexicalism. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '17 conference*, 200–229. Stanford, CA: CSLI Publications.
- Findlay, Jamie Y. 2019. *Multiword expressions and the lexicon*. Oxford: University of Oxford. (D.Phil. Thesis).
- Forst, Martin & Tracy Holloway King. 2021. Computational implementations and applications. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 141–180. Berlin: Language Science Press. DOI: ??.
- Fukui, Naoki. 1986. *A theory of category projection and its applications*. Cambridge, MA: Massachusetts Institute of Technology. (Doctoral dissertation).
- Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

## 2 Core concepts of LFG

- Hale, Ken. 1983. Warlpiri and the grammar of non-configurational languages. *Natural Language & Linguistic Theory* 1. 5–47. DOI: 10.1007/bf00210374.
- Halle, Morris & Alec Marantz. 1993. Distributed morphology and the pieces of inflection. In Kenneth Hale & Samuel Jay Keyser (eds.), *The view from Building 20: Essays in linguistics in honor of Sylvain Bromberger*, 111–176. Cambridge, MA: The MIT Press.
- Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80. DOI: 10.1515/flin-2017-1005.
- Haug, Dag & Tatiana Nikitina. 2015. Feature sharing in agreement. *Natural Language & Linguistic Theory* 34. 865–910. DOI: 10.1007/s11049-015-9321-9.
- Hoffmann, Thomas & Graeme Trousdale (eds.). 2013. *The Oxford handbook of Construction Grammar*. Oxford: Oxford University Press. DOI: 10.1093/oxfordhob/9780195396683.001.0001.
- Jackendoff, Ray. 1977. *X̄ syntax: A study of phrase structure* (Linguistic Inquiry Monographs 2). Cambridge, MA: The MIT Press.
- Jackendoff, Ray. 1992. Babe Ruth homered his way into the hearts of America. In Tim Stowell & Eric Wehrli (eds.), *Syntax and the lexicon* (Syntax and Semantics 26), 155–178. San Diego, CA: Academic Press. DOI: 10.1163/9789004373181.
- Kameyama, Megumi. 1985. *Zero anaphora: The case of Japanese*. Stanford, CA: Stanford University. (Doctoral dissertation).
- Kaplan, Ronald M. 1989. The formal architecture of Lexical-Functional Grammar. *Journal of Information Science and Engineering* 5. 305–322. Revised version published in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 7–27).
- Kaplan, Ronald M. & Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 173–281. Cambridge, MA: The MIT Press. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 29–130).
- Kaplan, Ronald M. & John T. Maxwell III. 1996. *LFG Grammar Writer's Workbench*. Xerox Palo Alto Research Center. Palo Alto, CA. [https://www.researchgate.net/profile/John\\_Maxwell5/publication/2760068\\_Grammar\\_Writer's\\_Workbench/links/0c96052405e97928e9000000.pdf](https://www.researchgate.net/profile/John_Maxwell5/publication/2760068_Grammar_Writer's_Workbench/links/0c96052405e97928e9000000.pdf).
- Kaplan, Ronald M. & Jürgen Wedekind. 1993. Restriction and correspondence-based translation. In *Proceedings of the 6th conference of the European chapter of the ACL (EACL 1993)*, 193–202. DOI: 10.3115/976744.976768.
- Kaplan, Ronald M. & Annie Zaenen. 1989a. Functional precedence and constituent structure. In Chu-Ren Huang & Keh-Jiann Chen (eds.), *Proceedings of ROCLING II*, 19–40. Taipei.



Oleg Belyaev

- Kaplan, Ronald M. & Annie Zaenen. 1989b. Long-distance dependencies, constituent structure, and functional uncertainty. In Mark Baltin & Anthony Kroch (eds.), *Alternative conceptions of phrase structure*, 17–42. Chicago: University of Chicago Press. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 137–165).
- Kharytonava, Volha (Olga). 2012. Taming affixes in Turkish: With or without residue? In Thomas Stolz, Hitomi Otsuka, Aina Urdze & Johan van der Auwera (eds.), *Irregularity in morphology (and beyond)* (Studia Typologica 11), 167–185. Berlin: De Gruyter.
- King, Tracy Holloway. 1997. Focus domains and information structure. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '97 conference*, 2–13. Stanford, CA: CSLI Publications.
- Kokkonidis, Miltiadis. 2007. Towards a more lexical and functional type-logical theory of grammar. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '07 conference*, 271–292. Stanford, CA: CSLI Publications.
- Kroeger, Paul R. 1993. *Phrase structure and grammatical relations in Tagalog*. Stanford, CA: CSLI Publications.
- Kuhn, Jonas. 2001. Resource sensitivity in the syntax-semantics interface: Evidence from the German split NP construction. In W. Detmar Meurers & Tibor Kiss (eds.), *Constraint-based approaches to Germanic syntax* (Studies in Constraint-Based Lexicalism), 177–216. Stanford, CA: CSLI Publications.
- Kuhn, Jonas. 2003. *Optimality-Theoretic Syntax – A declarative approach*. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2014. An LFG analysis of verbal modifiers in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 346–366. Stanford, CA: CSLI Publications.
- Langacker, Ronald W. 1969. On pronominalization and the chain of command. In David A. Reibel & Sanford A. Schane (eds.), *Modern studies in English*, 160–186. Englewood Cliffs, NJ: Prentice-Hall.
- Lovestrand, Joseph & John J. Lowe. 2017. Minimal c-structure: Rethinking projection in phrase structure. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '17 conference*, 285–305. Stanford, CA: CSLI Publications.
- Lowe, John J. 2011. R̥gvedic clitics and ‘prosodic’ movement. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '11 conference*, 360–380. Stanford, CA: CSLI Publications.
- Lowe, John J. 2016. English possessive ‘s’: Clitic and affix. *Natural Language & Linguistic Theory* 34. 157–195.
- Lyons, John. 1968. *Introduction to theoretical linguistics*. Cambridge. UK: Cambridge University Press. DOI: 10.1017/cbo9781139165570.



## 2 Core concepts of LFG

- Marcotte, Jean-Philippe. 2014. Syntactic categories in the correspondence architecture. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 408–428. Stanford, CA: CSLI Publications.
- Mohanan, K. P. 1982. Grammatical relations and clause structure in Malayalam. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 504–589. Cambridge, MA: The MIT Press.
- Muskens, Reinhard. 2001. Categorical Grammar and Lexical-Functional Grammar. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '01 conference*, 259–279. Stanford, CA: CSLI Publications.
- Nordlinger, Rachel. 1998. *Constructive case: Evidence from Australian languages*. Stanford, CA: CSLI Publications.
- Oehrle, Richard T. 1999. LFG as labeled deduction. In Mary Dalrymple (ed.), *Semantics and syntax in Lexical Functional Grammar: The resource logic approach* (Language, Speech, and Communication), 319–357. Cambridge, MA: The MIT Press.
- Panova, Anastasia. 2020. A case of morphologically bound complementation in Abaza: An LFG analysis. In Miriam Butt & Ida Toivonen (eds.), *Proceedings of the LFG '20 conference*, 289–306. Stanford, CA: CSLI Publications.
- Patejuk, Agnieszka & Adam Przepiórkowski. 2014. Control into selected conjuncts. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 448–460. Stanford, CA: CSLI Publications.
- Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press & CSLI Publications.
- Przepiórkowski, Adam & Agnieszka Patejuk. 2012. On case assignment and the coordination of unlikes: The limits of distributive features. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 479–489. Stanford, CA: CSLI Publications.
- Pullum, Geoffrey K. 2013. The central question in comparative syntactic metatheory. *Mind & Language* 28(4). 492–521. DOI: [10.1111/mila.12029](https://doi.org/10.1111/mila.12029).
- Sadler, Louisa & Doug Arnold. 1994. Prenominal adjectives and the phrasal/lexical distinction. *Journal of Linguistics* 30(1). 187–226. DOI: [10.1017 / S0022226700016224](https://doi.org/10.1017/S0022226700016224).
- Sells, Peter. 1994. Sub-phrasal syntax in Korean. *Language Research* 30(2). 351–386.
- Sells, Peter. 1995. Korean and Japanese morphology from a lexical perspective. *Linguistic Inquiry* 26(2). 277–325.
- Shieber, Stuart M. 1986. *An introduction to unification-based approaches to grammar* (CSLI Lecture Notes 4). Stanford, CA: CSLI Publications.

*Oleg Belyaev*

- Spencer, Andrew. 2005. Case in Hindi. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '05 conference*. Stanford, CA: CSLI Publications.
- Toivonen, Ida. 2003. *Non-projecting words: A case study of Swedish particles*. Dordrecht: Kluwer Academic Publishers.
- Vikner, Sten. 1995. *Verb movement and expletive subjects in the Germanic languages*. Oxford: Oxford University Press.
- Wescoat, Michael T. 2002. *On lexical sharing*. Stanford, CA: Stanford University. (Doctoral dissertation).
- Wescoat, Michael T. 2005. English nonsyllabic auxiliary contractions: An analysis in LFG with lexical sharing. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '05 conference*. Stanford, CA: CSLI Publications.
- Wescoat, Michael T. 2007. Preposition-determiner contractions: An analysis in Optimality-Theoretic Lexical-Functional Grammar with lexical sharing. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '07 conference*, 439–459. Stanford, CA: CSLI Publications.
- Wescoat, Michael T. 2009. Udi person markers and lexical integrity. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 604–622. Stanford, CA: CSLI Publications.
- Zaenen, Annie. 1983. On syntactic binding. *Linguistic Inquiry* 14(3). 469–504.

## **Part II**

# **Grammatical modules and interfaces**



# Chapter 3

## Prosody and its interfaces

Tina Bögel

University of Konstanz

LFG has always had a strong focus on syntax and semantics, but the last two decades have seen significant progress with regard to the integration of p(honological)-structure into LFG. This chapter first briefly introduces important concepts for the analysis of prosody and gives an overview of widely adopted approaches to the syntax–prosody interface. The second part surveys the different proposals for the integration of p-structure and its interfaces into LFG, with a particular focus on the architectural assumptions behind each approach and the resulting implications for the architecture of grammar.

### 1 Introduction

LFG has always had a strong focus on syntax and semantics. However, with the realisation that prosodic information can significantly contribute to linguistic analyses and is often crucial for the correct interpretation of meaning (e.g., in form of prosodic disambiguation of syntactically ambiguous structures or for the correct interpretation of information structure), the last two decades have seen significant progress with regard to the integration of prosodic structure into LFG.

LFG assumes that different aspects of grammar (i.e., syntax, semantics, etc.) are represented by unique modules (also called ‘projections’), each guided by its own principles and constraints, and with representations well-suited to their unique functions (cf. Dalrymple 2001; Sadock 1991, see also Belyaev 2021b [this volume]). The syntactic component, for example, is represented by c(onstituent)- and f(unctional)-structure and is concerned with constituency (via phrase structure rules) and the encoding of grammatical functions and morphosyntactic features, while phonology (including prosody) is represented by p(honological)-



*Tina Bögel*

structure and is concerned with phonological and prosodic properties like prosodic phrasing, rhythmic constraints, and intonation.<sup>1</sup>

Communication between the different modules is handled by LFG's correspondence architecture, which allows for relevant information to be made available at the respective interfaces. The establishment of these interfaces necessarily presumes a specific grammar architecture; that is, it presupposes an explicit positioning of modules with respect to each other. Discussing the architectural assumptions made in each p-structure proposal is thus essential for the understanding of the (in parts fundamental) differences in the representation of prosody and the communication at the interfaces.

This chapter provides an overview of the different approaches to prosody and its interfaces in LFG, and places these with respect to proposals made in the wider literature. It furthermore discusses the architectural assumptions made in each proposal and offers insights into a more general view of grammar. The chapter is structured as follows: Section 2 provides a general introduction into two major aspects of prosody (phrasing and intonation) and discusses current approaches to prosody and its interfaces in the wider literature. A discussion of the LFG grammar architecture and the placement of the phonological module (including prosody) is provided in Section 3. This section also establishes a fundamental difference between the proposals with respect to how grammar is viewed in general. Section 4 provides a chronological overview of the different approaches to prosody and its interfaces in LFG, in particular with respect to the architectural assumptions made in each proposal. Section 5 concludes the chapter.

## 2 Prosody and its interfaces

The LFG approaches to prosody discussed in Section 4 draw on a number of notions and theories established in the wider literature. This section first gives an overview on the general features that are particularly relevant with respect to the analysis of prosody at the interfaces and then describes the major approaches to the interface between syntax and prosody.

---

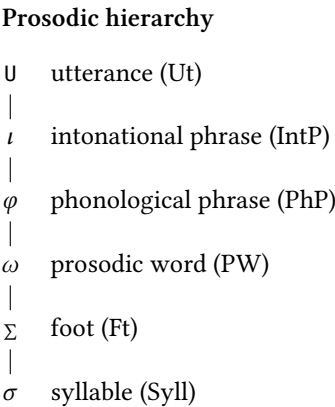
<sup>1</sup>In most of the approaches discussed in Section 4 the 'p' in p-structure represents p(rosodic)-structure, as prosodic features usually contain relevant information for analyses at the interfaces to syntax, semantics, and information structure. However, prosody is only one part of the larger field of phonology and some phenomena that are not part of prosody (e.g., postlexical sandhi phenomena) can be closely interlaced with prosody in that they can indicate a specific prosodic domain. This chapter will thus use the term p(honological)-structure, of which prosody is part, but which does not, per se, restrict p-structure to represent prosody alone.

2.1 What is prosody?

Prosody is a term used to describe suprasegmental phonology. It goes beyond the phonemic level of segmental phonology and is concerned with larger units in spoken language, including prosodic grouping, intonation and/or tones, rhythm, and stress patterns. Prosody can be used to express a number of properties and functions, among these clause type, clause structure, semantic scope, concepts of information structure such as topic and focus, but also speaker emotion, irony, or sarcasm. A detailed description of prosody goes far beyond the aim of this chapter and this section will only focus on some basic notions of prosody deemed fundamental for the current state of the art in LFG, namely prosodic phrasing, intonation, and the relationship between prosody and other modules of grammar.

Traditionally, it is assumed that spoken language is grouped into hierarchically structured **prosodic domains** (e.g., Selkirk 1978; Nespor & Vogel 1986; Hayes 1989). Example (1) shows the the most widely used proposal for the prosodic hierarchy originally made in Selkirk (1978) (building on an earlier proposal by McCawley 1968; see also Frota 2012 for different suggestions).

- (1) The Prosodic Hierarchy (Selkirk 1978) and its restrictions (Selkirk 1995: ex. 4)



In addition, the following constraints are assumed to apply to the prosodic hierarchy.<sup>2</sup>

- (2) **Constraints on Prosodic Domination**  
(where  $C^n$  = some prosodic category)

<sup>2</sup>These constraints have been challenged and are now mostly considered to be ‘soft’ constraints, see, e.g., Bennett & Elfner (2019).

Tina Bögel

- (i) *Layeredness* No  $C^i$  dominates a  $C^j$ ,  $j > i$ ,  
e.g. “No syllable dominates a foot.”
- (ii) *Headedness* Any  $C^i$  must dominate a  $C^{i-1}$  (except if  $C^i$  = syllable),  
e.g. “A prosodic word must dominate a foot.”
- (iii) *Exhaustivity* No  $C^i$  immediately dominates a constituent  $C^j$ ,  $j < i-1$ ,  
e.g. “No prosodic word immediately dominates a syllable.”
- (iv) *Nonrecursivity* No  $C^i$  dominates  $C^j$ ,  $j=i$ ,  
e.g. “No foot dominates a foot.”

The identification of a prosodic unit is based on various types of evidence and can vary greatly across languages. Among these types of evidence are sandhi processes (e.g., linking and intrusive /r/ in English (Wells 1970)), tonal events (e.g., Beckman & Pierrehumbert 1986; Pierrehumbert & Beckman 1988), and rhythmic patterns (e.g., Liberman 1975; Nespor & Vogel 1989). A phonological phrase in English, for example, is assumed to be intonationally represented by a pitch accent and a phrase accent, and to show phrase-final lengthening, where the last syllable is significantly longer compared to the other syllables of the phonological phrase (Lehiste et al. 1976; Frota 2012).

**Tonal events** like accents and boundary tones can contribute significantly to the meaning of a clause. These events are often described in terms of High and Low tones and tone combinations following the ToBI annotation conventions.<sup>3</sup> The first set of conventions was developed for American English in 1992 (Silverman et al. 1992); others have followed with specific adaptations to other languages (e.g., German GToBI (Grice & Baumann 2002)). The ToBI conventions distinguish between three tonal events:

- Pitch accents ( $L^*$  and  $H^*$ , and combinations like  $L+H^*$  and  $L^*+H$ ) are usually found on the words that are most important for an interpretation. In a neutrally pronounced sentence like *Amra went to the playground to meet her friends*, ‘Amra’, ‘playground’ and ‘friends’ would usually carry pitch accents. Pitch patterns can reflect information structure (**chapters/InformationStructure**): Contrastive focus in Germanic languages, for example, can be indicated by the use of an accent with a notably larger pitch span compared to the other accents of the clause (see, e.g., Féry 2020).

<sup>3</sup>The Autosegmental-metrical/Tone and break indices framework (AM/ToBI) (Pierrehumbert 1980; Silverman et al. 1992; Beckman et al. 2005) is a generally adopted set of conventions to describe tonal events in the intonational contour. Break indices, which indicate the strength of a break between words, are not further discussed in this chapter.



### 3 Prosody and its interfaces

- Boundary tones (H% and L%) are only associated with phrase edges of larger prosodic units, most often the intonational phrase boundary. They can, for example, signal the difference between a question and a statement with identical linear word order by means of rising or falling final phrase boundary tones.
- Phrase accents (H- and L-) are situated between a pitch accent and a boundary tone. They are often related to the edge of a prosodic domain below the intonational phrase, but there is some variation (see the discussion in Grice et al. 2000). They can significantly contribute to the disambiguation of syntactically ambiguous structures.

While these conventions are adopted by the vast majority of the field, proposals with a more fine-grained understanding of tonal events in combination with, for example, a distinct level of prominence (which is essential for the interpretation of focus type), have recently been developed (e.g., DIMA (Kügler et al. 2019)). Whether these proposals allow for a more thorough interpretation of prosody and meaning is subject to future research.

Both areas, prosodic constituency and intonation, are deeply intertwined with each other, and are also closely associated with **segmental phonology**, in that phonological processes (e.g., resyllabification) may be constrained to a particular prosodic domain (e.g., the phonological phrase), or the quality of a vowel may change if it is associated with a pitch accent. Segmental and suprasegmental prosody both are part of lexical and postlexical phonology. Prosodic constituency and (lexical) stress are also part of a word's lexical entry, as is the knowledge about prosodically deficient clitics, while segmental phenomena frequently also occur between two words and hence are not restricted to the lexicon. As a consequence, p-structure should not only represent prosodic structure, but should rather include lexical and postlexical segmental and suprasegmental phonology (cf. fn 1).

**Phonetics** can be viewed as the physical translation of phonology into a concrete speech signal (and vice versa), which is reflected in the close relationship between prosodic terms like pitch, length, or loudness, and phonetic terms like fundamental frequency, duration, and intensity (see Kingston 2019 for a detailed discussion). Phonetics has not been in the focus of the proposals made in LFG, although initial approaches towards its integration have been undertaken (see Butt et al. 2020; Bögel 2020; also Section 4.5).

It is clear that prosodic structure is governed by p-structure internal principles and constraints, and that, for example, rhythm and the prosodic status of words

*Tina Bögel*

(prosodic words vs. prosodically deficient clitics) can determine the formation of prosodic domains. It is, however, equally assumed that prosody is influenced by syntactic structure and discourse-related aspects like the differentiation between new and given information and the expression of different focus-types (**chapters/InformationStructure**). Furthermore, ‘extralinguistic’ factors such as speaking rate or frequency effects can affect p-structure.<sup>4</sup> The exact influence of these and other factors on prosody is far from being fully explored. As the vast majority of research (within and outside of LFG) has focussed on the exploration of the relationship between syntax and prosody, the major approaches to this interface are briefly introduced here, before turning to the role of p-structure and the different proposals to prosody and its interfaces in LFG.

## 2.2 Theories of the prosody-syntax interface

The literature on how the syntactic and the prosodic modules interact can be roughly divided into two major camps: **direct reference** and **indirect reference** (see Bennett & Elfner 2019 for a detailed discussion of each approach). The direct reference approach proposes that phonological rules and groupings can directly be conditioned on syntactic relations or properties, e.g., on c-command, sister relations, or ‘head’ status, without the intervention of a separate prosodic structure (e.g., Kaisse 1985; see Elordieta 2008 for an overview). As LFG assumes a modular view of grammar and none of the LFG approaches propose the (non-modular) direct reference approach, this chapter will not further discuss this particular approach to the interface.

The other school of thought pursues the indirect reference approach, which assumes that syntactic structure is first mapped to prosodic domains as shown in (1). Phonological rules are then conditioned based on these prosodic domains (e.g., Hayes & Lahiri 1991). Prominent proposals include the **end-based** approach (Selkirk 1986; Chen 1987) which assumes that the mapping algorithm is restricted to the edges of syntactic heads and maximal projections. In the following abstract example, each syntactic head receives a prosodic word boundary and each XP receives a phonological phrase boundary at its right edge. As all XPs align at their right edge, only one phonological phrase boundary is included.<sup>5</sup> Function words (‘fw’) are excluded from the mapping algorithm.

---

<sup>4</sup>See Shattuck-Hufnagel & Turk (1996) for a thorough discussion of different constraints on prosody.

<sup>5</sup>Whether the right or the left edge is aligned seems to be subject to language-specific constraints.

### 3 Prosody and its interfaces

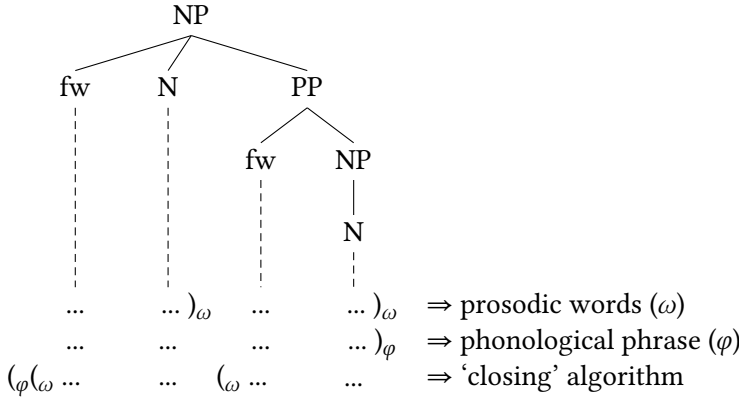


Figure 1: The end-based approach (Selkirk 1986: 387, shortened and modified)

All prosodic domains in this example are ‘closed’ by assuming an automatic insertion of a left boundary with the neighbouring right boundaries or at the edges of the whole construction. This effectively groups all function words together with their corresponding syntactic heads into prosodic words and places both prosodic words within the phonological phrase:  $(_{\phi}(_{\omega}\text{fw N})_{\omega}(_{\omega}\text{fw N})_{\omega})_{\phi}$ .

The end-based approach has been reformulated as a generalized alignment constraint in Optimality Theory (OT; McCarthy & Prince 1993; Prince & Smolensky 2004) and is generally represented in this format, for example as ALIGN-XP-R(IGHT): ALIGN (XP, R;  $\phi$ , R)<sup>6</sup> (Selkirk 1995).

In later work, Selkirk (2009, 2011) introduced **match theory** (which has ancestors in, e.g., Ladd 1986). In contrast to the previous end-based approach, match theory assumes that both edges of a syntactic constituent are simultaneously matched to a prosodic constituent.

- **MATCH-CLAUSE:** A clause in syntactic constituent structure must be matched by a corresponding intonational phrase ( $\iota$ ) in prosodic constituent structure: MATCH (CLAUSE,  $\iota$ )
- **MATCH-PHASE:** A phrase in syntactic constituent structure must be matched by a corresponding phonological phrase ( $\phi$ ) in prosodic constituent structure: MATCH (XP,  $\phi$ )

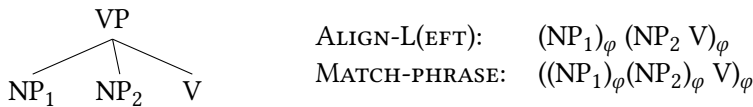
<sup>6</sup>Read as: “Align the right edge of an XP with the right edge of a phonological phrase/ $\phi$ ”.

Tina Bögel

- MATCH-WORD: A (lexical) word in syntactic constituent structure must be matched by a corresponding prosodic word ( $\omega$ ) in prosodic constituent structure: MATCH (LEX-WORD,  $\omega$ ) (Selkirk 2011: 439, modified)

Match theory reflects the syntactic structure in much more detail; in particular (and in contrast to the end-based approach) it predicts recursion, as syntactically nested XPs will be phrased as recursive structures in prosodic constituency. The (modified) example in (3) from Selkirk (2011) illustrates this point for a transitive verb phrase.

- (3) Different mapping approaches for a transitive verb phrase



In (3), each VP/NP receives a preceding left phonological phrase boundary in the end-based approach (ALIGN-L). The  $\varphi$ -boundary for  $\text{NP}_1$  is identical with the boundary for the VP. The right boundary for  $\text{NP}_1$  is placed ('closed') before  $\text{NP}_2$ , and the second right  $\varphi$ -boundary is placed after V, which does not receive any boundaries by itself. The first phonological phrase thus contains  $\text{NP}_1$  and the second phonological phrase groups the verb together with  $\text{NP}_2$ . In contrast, the MATCH algorithm maps each XP ( $\text{NP}_1$ ,  $\text{NP}_2$ , and VP) into a phonological phrase, thus creating a recursive structure.

Besides these two major schools of direct and indirect reference, a third proposal with regard to the formation of prosodic structure has been adopted into the LFG community as well (most prominently in Dalrymple & Mycock 2011; Mycock & Lowe 2013, and subsequent work), which will be called the **parallel approach** in this chapter. The main motivation for the parallel approach is the frequently observed non-isomorphism between syntactic and prosodic constituency as illustrated in the following example.

- (4) Syntactic Phrasing:    [Drink [[a pint] [of milk]] [a day]]  
 Phonological Phrasing:    (Drink a) (pint a) (milk a) (day)  
 (Lahiri & Plank 2010: 376, modified)

This frequent mismatch seemingly rules out any approaches which map syntactic constituents to prosodic domains, but suggests that prosodic structure is built up on prosodic principles alone. Based on observations of rhythmic patterns (e.g., Sweet 1904), Lahiri & Plank (2010) assume the trochaic foot (X –) to be the determining element for the creation of prosodic structure in English,

### 3 Prosody and its interfaces

with the stressed syllable as the initial element of each prosodic chunk. Lahiri and Plank discuss this approach to prosodic phrasing with respect to a number of diachronic and synchronic examples which support the assumption that function words are frequently grouped together with preceding strong syllables (as in example 4) and not necessarily with the following syntactic head (as suggested in Figure 1).

However, while the parallel approach provides a suggestion for the creation of the lower prosodic domains (foot and prosodic word structure), no suggestion is made for the formation of the higher domains (phonological phrase and intonational phrase); nor do Plank and Lahiri explicitly exclude the influence from other modules of grammar. Furthermore, it has long been part of the indirect reference tradition that it is “crucially [...] not the case that all syntactic boundaries of a certain type must correspond to prosodic boundaries of a given type and vice versa” (Frota 2012: 256). Most researchers assume “prosodic restructuring” (Nespor & Vogel 1986: 172) based on, for example, the type of word (function word vs. lexical word), the size of the phonological phrase, or the amount of recursive nesting.<sup>7</sup> The indirect reference approaches thus do not take prosodic constituency to be a simple derivative of syntax, but assume that prosodic structure is also formed by means of syntax-independent constraints, among them the rhythmic constraints proposed by the parallel approach.

The difference between these two approaches to prosody and its interfaces necessarily reflects two distinct views of grammar in general. As both directions have been pursued in the proposals made in LFG, the following section briefly discusses how prosodic structure is integrated into the overall grammar architecture and how the indirect reference approach and the parallel approach differ with respect to the communication at the interfaces.

### 3 Prosody in LFG’s grammar architecture

Several proposals have been made with respect to LFG’s grammar architecture (see also Belyaev 2021b [this volume]) and a closer discussion of the different approaches to prosody and its interfaces presented in this chapter provides interesting insights into the positioning of the different modules on the one hand, and a general understanding of grammar on the other hand.

Prosodic structure is especially interesting as it is usually taken to represent FORM and is thus placed at one ‘end’ of the FORM-MEANING relationship as, for ex-

---

<sup>7</sup>For prosodic restructuring mechanisms/‘prosodic markedness constraints’ as expressed in Optimality Theory, see Selkirk (2011: 468ff) for an overview.

Tina Bögel

ample, discussed in Kaplan (1987: 362). In Kaplan’s original proposal, p-structure was not yet part of the grammar architecture; instead, the (word) string was taken to be the external form of the sentence. This is also reflected in Asudeh (2009: 110), where the string is understood as a ‘linear representation of phonology’ and as the center of the syntax–phonology interface, a proposal that has concretely been pursued in the approaches of Bögel et al. (2009, 2010), Dalrymple & Mycock (2011), and Mycock & Lowe (2013) (see Section 4).

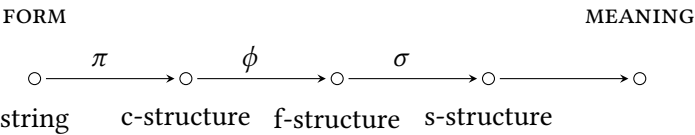


Figure 2: (Simplified) FORM-MEANING relationship in Kaplan (1987: 362)

Proposals in the wider literature have assumed a slightly more complex representation of FORM. Very early, Selkirk (1984) proposed that syntactic structure is first mapped to a phonological (including prosodic) representation, which is then further processed by means of phonological rules and constraints before being mapped to a phonetic representation. In this model the string is not placed between the syntactic and the phonological module, but is the output of the phonological and the phonetic modules. Such a model is very much in line with psycholinguistic models of speech production and comprehension, e.g., as found in Levelt (1999) and in Jackendoff (2002)’s work on Parallel Grammar; see **chapters/SimplerSyntax**.

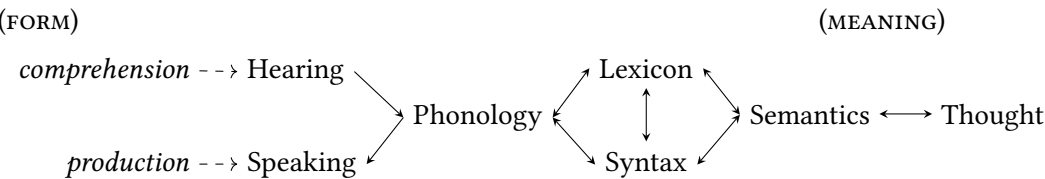


Figure 3: The language processor (cf. Jackendoff 2002: 197, modified)

The model in Figure 3 states clearly what is only implicitly expressed in theoretical LFG:<sup>8</sup> The different modules, placed between FORM and MEANING, assume

<sup>8</sup>For example, in Figure 2 by means of arrows, and more explicitly in the pipeline architectures of the numerous computational LFG grammar implementations (see Forst & King 2021 [this volume]).

### 3 Prosody and its interfaces

a certain directionality, generally termed as ‘comprehension’ (parsing) and ‘production’ (generation) in the wider literature. This is seemingly in conflict with the assumption that the different modules exist in ‘parallel’ in LFG (Dalrymple et al. 2019: 265); however, as Jackendoff explicitly remarks, this is not necessarily a hindrance:

P[arallel] A[rchitecture] is nondirectional, but its constraints can be implemented in any order suited to particular processing tasks. (Jackendoff 2010: 589)

‘Parallel’ under this approach refers to the general understanding that each module is subject to its own principles and constraints (= modularity). It does not mean that each component builds a completely isolated structure which then has to be aligned to the output of other modules. Instead, the individual constraints should be adjusted to the processing task at hand (which is either from FORM to MEANING (comprehension/parsing) or from MEANING to FORM (production/generation)).

While this distinction might not carry much weight if a linguistic analysis is provided within one module of grammar (e.g., a syntactic phenomenon), it is crucial when modelling constraints at an interface, as the involvement of two (or more) modules always involves a ‘direction’. The assumptions made by Selkirk above, for example, are made from the perspective of production, while the architecture proposed by Kaplan in Figure 2 (and in general the vast majority of LFG-related linguistic analyses) is made from the comprehension perspective. The acknowledgement of this bidirectionality as made explicit in Figure 3 is fundamental for the discussion of any interfaces between different modules, and thus essential for the proposals on the integration of prosody and its interfaces into LFG.

Models which follow the parallel approach to prosody and its interfaces as detailed in Section 2.2 by assuming that modules are built up independently of each other and that their output is matched for the best alignment at each interface might seemingly be in line with the concept of modules existing in parallel. However, such models are not built to reflect the processing of a given speech signal to understand its meaning (→ comprehension), or the production of a signal expressing a specific thought (→ production).

## 4 LFG approaches to prosody and its interfaces

With respect to prosody and its interfaces, both the indirect reference and the parallel approach have been explored within the LFG community, mostly with a

Tina Bögel

directional perspective. As these proposals frequently influence each other and furthermore represent very different views of grammar, the following section provides a chronological overview of the different approaches with a specific focus on the architectural assumptions behind each proposal.

### 4.1 From c-structure to p-structure: Butt & King (1998)

Butt & King (1998) were the first to introduce a discussion of the syntax-prosody interface and p-structure to LFG. They assumed a mutually constraining model where d(iscourse)- and p-structure are projected off c-structure (in parallel to f-structure).

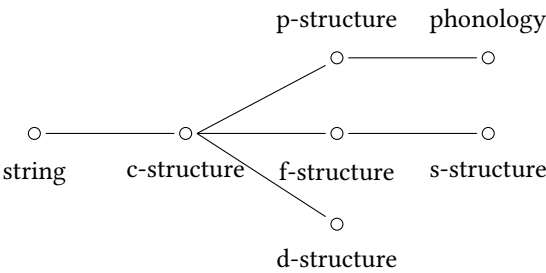


Figure 4: Grammar architecture according to Butt & King (1998: modified)

Under this approach, c-structure is a pivot point between d- and p-structure.<sup>9</sup> P(rosodic)-structure is viewed as an intermediate between c-structure and the phonological component itself which also contains postlexical phonological rules.

Based on work by Hayes & Lahiri (1991), Butt and King focus on syntactically ambiguous sentences in Bengali.

- (5) ami bHut dekH-I-am  
I ghost see-PST-1SG
- a. ‘I was startled’ (idiomatic)
- b. ‘I saw a ghost.’ (transitive)

Following findings discussed in Hayes and Lahiri, Butt and King assume that prosody can be applied to differentiate between the idiomatic and the transitive interpretation. For Bengali, the assumption for the syntactic-prosodic constituent mapping is that every clause is mapped to an Intonational Phrase (ι,

<sup>9</sup>In contrast to this chapter which takes prosody to be part of phonology (see footnote 1), Butt and King differentiate between a p(rosodic)-structure and a phonological component.



### 3 Prosody and its interfaces

IntP), every NP to a Phonological Phrase ( $\varphi$ , PhP), and every main V or complex predicate is phrased separately. Figure 5 shows the c-structures for (5) and the phrasing possibilities for *bHut dekHlam* which consist of separate phonological phrases in the transitive, and a single phonological phrase in the idiomatic reading.

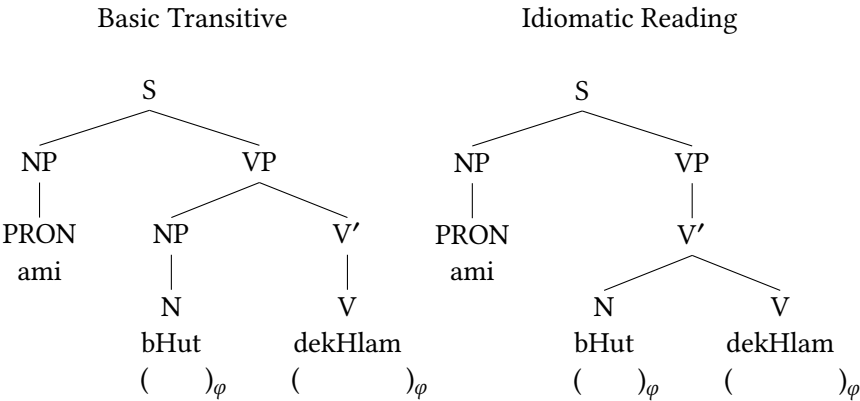


Figure 5: Two c-structure analyses for example (5), Butt and King (1998, modified)

Butt & King (1998) represent p-structure as an AVM structure based on the prosodic hierarchy as shown in (1) above. The AVM structure allows for the inclusion of more detailed information beyond the prosodic domain and the p(honological)-form, such as pitch accents or boundary tones. Butt and King also discuss the linearity issue given with any AVM approach. In order to apply phonological and phonetic processes, it is necessary to preserve the linear order of the string. For a possible solution to this issue, Butt and King point towards projection precedence (Zaenen & Kaplan 1995), which arranges the attributes in p-structure similarly to the string.

Figure 6 shows the prosodic structure of the idiomatic reading in (5), where *bHut* and *dekHlam* are phrased into one phonological phrase (DOM(AIN): P-PHRASE). The AVM includes all the information in the tree structure and the additional information known about the tones in the language, for example, that in a neutral (non-phonological) focus construction, a high tone is associated with the left p-word in the rightmost p-phrase and the whole clause receives a low boundary tone. This information is stored in the AVM (TONE HIGH), but the final association of the hight tone with the correct p-form (and the correct syllable in this p-form) is left to the phonological component itself. The reason is that the final p-phrase can only be identified once prosodic phrasing is complete and that the placement

Tina Bögel

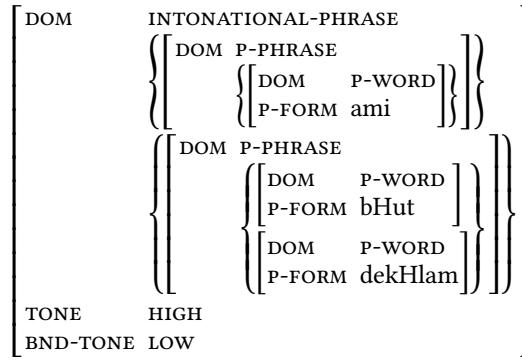


Figure 6: Prosodic structure relating to the idiomatic reading in example (5), neutral focus

of the pitch accent on the correct syllable is solely depending on the phonological structure of the word itself.

Contrastive focus, on the other hand, is indicated by a low pitch accent and a high (intermediate) boundary tone at the level of the phonological phrase. As the target of the contrastive focus is determined by grammar (here d(iscourse)-structure), the associated pitch accent and boundary tone can be mapped to p-structure together with their domain.

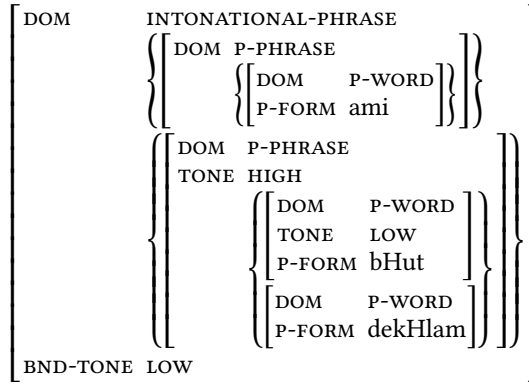


Figure 7: Prosodic structure relating to the idiomatic reading in example (5), contrastive focus

Butt and King determine the tone distribution by using c-structure as a pivot

## 3 Prosody and its interfaces

between d-structure and p-structure.<sup>10</sup>

- (6)  $(\downarrow_d \text{ FOCUS-TYPE}) =_c \text{ CONTRASTIVE}$   
 $(\uparrow_p \text{ TONE}) = \text{ HIGH} \rightarrow \text{ phrasal high}$   
 $(\downarrow_p \text{ TONE}) = \text{ LOW} \rightarrow \text{ local low}$

The approach proposed by Butt and King (1998) was later taken up in Bögel et al. (2008) in their analysis of Urdu *ezafe*. In the *ezafe* construction in (7), the *ezafe* clitic is syntactically grouped with the following modifying noun, but is prosodically attached to the previous head noun.

- (7) sher=e panjAb  
 lion=Ez Punjab  
 ‘a/the lion of Punjab’  
 Syntactic Phrasing: [[sher] [e panjAb]]  
 Prosodic Phrasing: ((sher e) panjAb)

Example (7) shows a typical mismatch between syntactic and prosodic structure, which is difficult to account for if prosodic constituency is directly based on syntactic constituency. The solution proposed in Bögel et al. (2008) integrates the *ezafe* clitic (CL-FORM) into the phonological phrase using a number of bookkeeping features to make sure an *ezafe* clitic is present, as shown in Figure 8.

$$\begin{array}{ll}
 \text{NPez}' \rightarrow \text{N:} & \begin{array}{l} (\uparrow_p \text{ DOM}) = \text{P-WORD} \\ (\downarrow_p \text{ P-FORM}) = \text{sher} \\ (\downarrow_p \text{ CL-FORM}) = \text{ezafe} \\ (\uparrow \text{ CHECK EZAFA}) =_c + \end{array} \\
 \text{ezafe:} & (\uparrow \text{ CHECK EZAFA}) = +
 \end{array}
 \left[ \begin{array}{l} \text{DOM P-PHRASE} \\ \left\{ \left[ \begin{array}{l} \text{DOM P-WORD} \\ \left\{ \left[ \begin{array}{l} \text{P-FORM sher} \\ \text{CL-FORM EZAFA} \end{array} \right] \right\} \right] \right\} \right\} \\ \left\{ \left[ \begin{array}{l} \text{DOM P-WORD} \\ \left\{ \left[ \text{P-FORM panjAb} \right] \right\} \right] \right\} \right\} \end{array} \right]$$

Figure 8: (Reduced) *ezafe* rule and the resulting p-structure in Bögel et al. (2008)

This approach is not entirely satisfactory. For one, in this approach it is actually the noun which is ‘checking’ for a following clitic, instead of the clitic ‘asking’ to be grouped with a preceding prosodic host. Furthermore, this approach does not allow for a language-specific expression of prosodic principles, e.g., the general integration of *enclitics* into the preceding prosodic domain, and of

<sup>10</sup>For the interested reader: Focus in Bengali can also be signalled by the clitic -o. Following Lahiri & Fitzpatrick-Cole (1999), Butt and King assume a lexical high tone which is introduced onto the prosodic word with the clitic’s lexical specifications.

Tina Bögel

*proclitics* into the following prosodic domain. Instead, individual specifications have to be created for each clitic. This is not only unintuitive, but also does not allow for any predictions to be made about prosodic structure in general.

Summing up, Butt and King make a first proposal to include prosodic information into LFG and show how this can interact with d- and c-structure. In contrast to Hayes & Lahiri (1991)'s original approach (and in contrast to the claim made in Dalrymple et al. 2019), their model does not permit the direct reference of phonological restructuring rules to relations internal to syntactic structure (e.g., to 'right sister' or modifier-head-constructions), but provides an indirect, modular approach to the interface.

Butt and King distinguish between two structures: p-structure and the phonological component. P-structure only includes the information that is pre-determined by other modules of grammar, e.g., pitch patterns introduced by different sentence types and focus, and prosodic constituency based on syntactic constituency. This information serves as input to the phonological component (not further defined in their paper) and its inherent rules and constraints, which include prosodic restructuring, or the placement of pitch accents on the correct syllables within the right domains. This directional analysis from c-structure to p-structure to phonology reflects part of the production process in Figure 3. However, Butt and King's model (as shown in Figure 4) is generally not in line with the architectural assumptions made in Figure 2 and Figure 3 in that the string is not the representative of FORM: neither is the string equal to the phonological output nor is it closely associated with the phonological module. Without this connection, it is unclear how the string could be realised in terms of a (physical) speech signal.

## 4.2 Prosody and i-structure: O'Connor (2005)

In his thesis, O'Connor (2005) discusses the interface between prosody and information structure. O'Connor (2005) assumes a bidirectional approach, which distinguishes between a 'hearer-based' and a 'speaker-based' approach. He explicitly only focusses on the hearer-based direction (from p- to d-structure → comprehension) and leaves the speaker-based direction (from d- to p-structure → production) to further research.

O'Connor's approach is based on the AM/ToBI framework (see Section 2.1), but the description of accents is restricted to High and Low tones only.<sup>11</sup> He is

---

<sup>11</sup>O'Connor does not distinguish between different types of pitch accents and how these may relate to specific i-structure categories, e.g., the distinction between broad and narrow focus based on different pitch patterns (a.o., Baumann et al. 2007).

## 3 Prosody and its interfaces

mostly concerned with utterances where a difference in meaning is expressed solely by means of prosodic emphasis (expressed by capital letters in example (8)).

- (8) a. He rode [a green DRAGON]<sub>loc</sub>.  
 b. He rode a [GREEN]<sub>loc</sub> dragon.

The two propositions have a different information structure. Example (8a) can be the answer for a question with a broad focus (e.g., *Did he ride a green dragon or a thestral?*), while example (8b) is more likely to be the (contrastive) answer to a question like *Did he ride a green dragon or a blue dragon?*

In his proposal, O'Connor assigns a central role to i-structure. As the AM/ToBI system is not concerned with the influence of syntactic structure on prosody, O'Connor assumes that prosody and i-structure can be related to each other without syntactic mediation.<sup>12</sup>

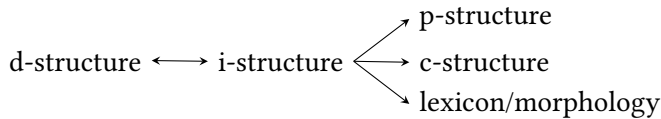


Figure 9: Architecture proposed in O'Connor (2005: Fig 6.3, p. 142, modified)

Following the general idea behind autosegmental approaches (Goldsmith 1976), O'Connor pursues the idea of a representation of tonal information independent from the segmental/phonemic representation. He proposes that p-structure should be represented by a hierarchical constituent structure (thus paralleling c-structure). Via so-called 'tune structure rules', a tree-like structure is created to represent intonation where the terminal nodes correspond to underspecified tonal events:  $t^*$  represents a pitch accent,  $t^-$  a phrase accent, and  $t\%$  a boundary tone.

- (9)  $n \geq 1$   
 a.  $TUNE_{IP} \rightarrow tune_{ip}^n t\%$   
 b.  $tune_{ip} \rightarrow t^{*n} t^-$

<sup>12</sup>O'Connor does not completely exclude the influence of c-structure on prosody, but only acknowledges a relevance of the linear and hierarchical syntactic structure of the clause for the length of the transition between tonal events and the alignment of the pitch in general.

Tina Bögel

As a result, each prosodic tree constructed on the basis of these rules has four obligatory nodes: the prosodic ‘intermediate phrase’ ( $\text{tune}_{ip}$ ) consists of a nuclear accent  $t^*$  and a phrase accent  $t^-$  while the prosodic ‘intonational phrase’ ( $\text{TUNE}_{IP}$ ) consists of at least one intermediate phrase and a boundary accent  $t\%$ . For example (5) (see also Figure 5, basic transitive) from Butt & King (1998), O’Connor proposes the p-structure in Figure 10.

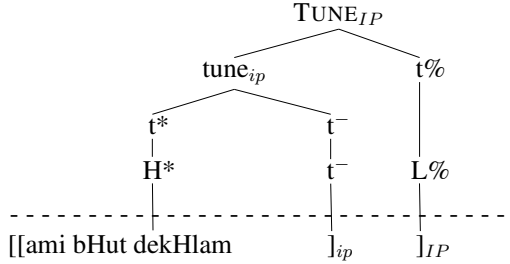


Figure 10: O’Connor’s p-structure as applied to example (5) from Butt & King (1998)

The tree representation is mainly concerned with the organisation of tonal events; the material below the dashed line includes the orthographic tier and the prosodic domain information in form of bracketing.<sup>13</sup>

O’Connor emphasises the point that under his approach, the association of the High tone is not left to a further phonological component as proposed in Butt and King (1998) and discussed above in Section 4.1. It is however not quite clear how the High tone is associated with the correct string sequence in O’Connor’s approach, as no formal alignment of string and pitch (i.e., c-structure and p-structure) is established in his thesis. Indeed, in the data provided by Butt and King, the High tone should be assigned to the ‘leftmost’ prosodic word in the ‘rightmost’ phonological phrase. As O’Connor collapses all three phonological phrases proposed by Butt and King under one  $\text{tune}_{ip}$ , it is not clear how the association of the High tone with the correct word can be ensured.

With respect to i-structure, O’Connor assumes that categories like FOCUS and TOPIC are organised linearly in an utterance, and assigns each to one  $\text{tune}_{ip}$ . If there are not enough tunes, then the assumption is that there is no topic correspondent.

$$(10) \quad \text{TUNE}_{IP} \longrightarrow \begin{array}{ccc} \text{tune}_{ip} & & \text{tune}_{ip} & t\% \\ \downarrow \in \{\uparrow_d \text{ TOPIC}\} & & \downarrow \in \{\uparrow_d \text{ FOCUS}\} & \end{array}$$

<sup>13</sup> As O’Connor’s (2005) main focus is on the relation between intonation and discourse functions, the encoding of further prosodic/phonological information, e.g., syllable structure, or lexical stress, is not further discussed in his thesis.

### 3 Prosody and its interfaces

As O'Connor notes, there are a number of cases where the proposed association of i-structure categories and tunes does not work. Sentences like *'It broke.'* will only have one tune, indicating a FOCUS. However, in i-structural terms, *It* is a TOPIC. This mismatch between tunes and i-structure roles is discussed (O'Connor 2005: 161), but not resolved.

In conclusion, O'Connor's indirect, directional approach to the relationship between prosody and i-structure suggests an alternative to the syntactocentric view proposed in Butt and King (1998). However, there are a number of outstanding questions. Besides the incomplete association of i-structure categories to tune-structure rules just discussed, there is also the fundamentally important question as to how tonal events can be associated with their targets without reference to the morphosyntactic string, or how common phenomena, e.g., the (syntactic) scope of a prosodically expressed focus, can be determined without reference to syntactic constituency. The missing association of p-structure, string, and c-structure, and other unresolved questions thus only allow for an analysis of a more descriptive nature.

#### 4.3 The string as an interface between c- and p-structure: Bögel et al. (2009, 2010)

Based on the realisation that the frequent misalignment of prosodic and syntactic constituents would seriously complicate previously established prosody-syntax mapping algorithms, Bögel et al. (2009) pick up on the notion of the parallel approach discussed in Section 2.2. The underlying assumption is that the prosodic component operates independently of syntax and that the two components are not related via LFG's projection architecture. To account for the cases where syntax is influenced by prosody, Bögel et al. (2009) assume a directional 'pipeline' architecture (from the comprehension perspective): First, an independent prosodic component interprets various phonological properties thus establishing the boundaries of prosodic units. This information on prosodic constituency is then made available to syntax by inserting prosodic bracketing features into the terminal string of c-structure.<sup>14</sup>

Bögel et al. (2009) discuss a number of different phenomena, among them Urdu *ezafe* (Bögel et al. 2008: see Section 4.1). They extend the c-structure rules by adding left and right prosodic brackets ('lexical categories' RB and LB) which reflect prosodic constituency.

---

<sup>14</sup>Under this approach, the string has a central role as it includes information from both the syntactic and the prosodic component (similar to the understanding of the string in Asudeh 2009: 110 as a 'linear representation of phonology').

Tina Bögel

- (11) a. EzP  $\rightarrow$  EZ RB N  
 b. NPez'  $\rightarrow$  LB [...] N

The inclusion of brackets greatly simplifies the rule originally used in Bögel et al. (2008: Figure 10) where a number of CHECK-features were applied to control for an *ezafe* clitic following the head noun. The resulting c-structure representation allows for the depiction of the misalignment between syntactic and prosodic structure.

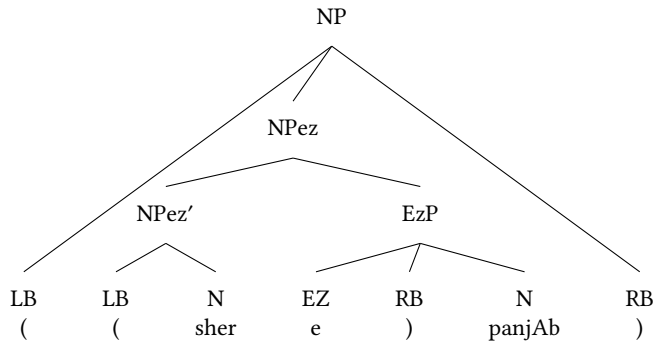


Figure 11: Urdu *ezafe* analysis as proposed in Bögel et al. (2009)

Another aspect discussed in Bögel et al. (2009) is the prosodic resolution of syntactically ambiguous structures. Consider the following example, where *old* can either modify only the first noun ((12a)) or scope over the whole coordination ((12b)). Each possibility is accompanied by a distinct prosodic grouping.

- (12) a. [old men] and [women]  
 (old men) and (women)  
 b. [old [men and women]]  
 (old (men and women))

The paper postulates a ‘Principle of Prosodic Preference’, according to which the syntactic component disprefers syntactic structures whose constituent boundaries do not coincide with prosodic boundaries. For the implementation, Bögel et al. (2009) use a metarule, which systematically transforms the rules of the syntactic component. In the following metarule, CAT is a nonterminal category, and RHS denotes the regular language over categories which are annotated with co-describing constraints.



## 3 Prosody and its interfaces

(13) CAT  $\rightarrow$  RHS

In Bögel et al.'s metarule in (14), the top part of the rule will match a (recursive) sequence of CAT surrounded by prosodic brackets (LB and RB). The bottom part will match the RHS regular expression if all occurrences of LB or RB are ignored, thus preventing the inserted prosodic brackets from ruling out a valid syntactic analysis.<sup>15</sup>

(14) CAT  $\rightarrow$  LB CAT RB  
           | RHS / [ LB | RB ]  
                                 Disprefer

The Principle of Prosodic Preference is enforced via the 'Disprefer' optimality mark,<sup>16</sup> which assigns a dispreference mark to the construction every time the bottom part of the metarule in (14) is applied. In the case of several possible syntactic analyses (as it is the case with (12)), this extension effectively ranks the analyses: The top half of the rule only applies if the prosodic brackets match the syntactic structure, while the syntactic analyses with no matching prosodic brackets will be parsed by the bottom half of the rule, but will receive a 'Disprefer' mark. This allows for constructions with matching prosodic and syntactic brackets to be preferred, while constructions with non-matching brackets will only be valid if a preferred solution (with matching brackets) is not available.

This first approach to the interface was extended in Bögel et al. (2010) which discusses second position (2P) clitics in Russian and Serbian/Croatian/Bosnian (SCB). It is concerned with examples like (15), where a clitic cluster (CCL) disrupts the NP *Taj čovek*.

(15) [Taj *joj ga je* čovek] poklonio.  
       that her it AUX man presented  
       'That man presented her with it.' (Schütze 1994)

These clitics appear in the second position after a first prosodic word without regard to syntactic requirements ((16)).

(16) (((Taj)<sub>ω</sub> (joj ga je)<sub>cl</sub>)<sub>ω</sub> (čovek)<sub>ω</sub>)<sub>φ</sub> (poklonio)<sub>φ</sub>  
       That her it AUX man presented  
       'That man presented her with it.'

<sup>15</sup>The 'Ignore operator' / was first introduced in Kaplan & Kay (1994).

<sup>16</sup>See Forst & King 2021 [this volume], Section 1.5 for a description of optimality marks and relevant references.

Tina Bögel

Such a structure is problematic for traditional LFG accounts, because it is difficult to account for the clitics' appearance within a NP and to furthermore retrieve the clitics' functional contribution to the clausal f-structure. With a purely syntactic account, this information is locked into the NP's f-structure.

Bögel et al. (2010) resolve this issue by assuming a shared responsibility between the syntactic and the prosodic component: While the syntactic component ensures the availability of the functional information by placing the clitic in the (linear) first position, the prosodic component ensures the correct position of the clitics within the clause and places the clitics following the first prosodic word. This prosodic repair mechanism has been shown to apply crosslinguistically and was dubbed 'prosodic inversion' by Halpern (1995).<sup>17</sup>

In order for the clitic to appear in the correct syntactic position, Bögel et al. (2010) define the following rule, where  $RHS_S$  denotes the possible expansion of the clausal S node with left and right brackets (as discussed above).  $LB_S$  is a pre-terminal node that marks the left edge of a clause and allows syntactic/prosodic constraints to be aligned with respect to clause boundaries. CCL can optionally appear as a prefix to the S expansion; the  $\uparrow=\downarrow$  annotation ensures the processing of the clitics' clause-level functional information.

$$(17) \quad S \longrightarrow LB_S \underset{\uparrow=\downarrow}{(CCL)} RHS_S$$

To account for the prosodic placement, Bögel et al. (2010) distinguish between a prosodic and a syntactic (c-structure terminal) string which includes the lexical formatives discussed above. The interface between these (usually aligned) strings is a regular relation, where the syntactic string is the 'upper language' and the prosodic string is represented by the 'lower language'. In the simplified illustration in (18), the upper language/syntactic string clitic sequence (CS:0) immediately following the clause boundary ( $_S$  is placed after the first prosodic word  $\omega$  in the prosodic string in the lower language/prosodic string (0:CS).

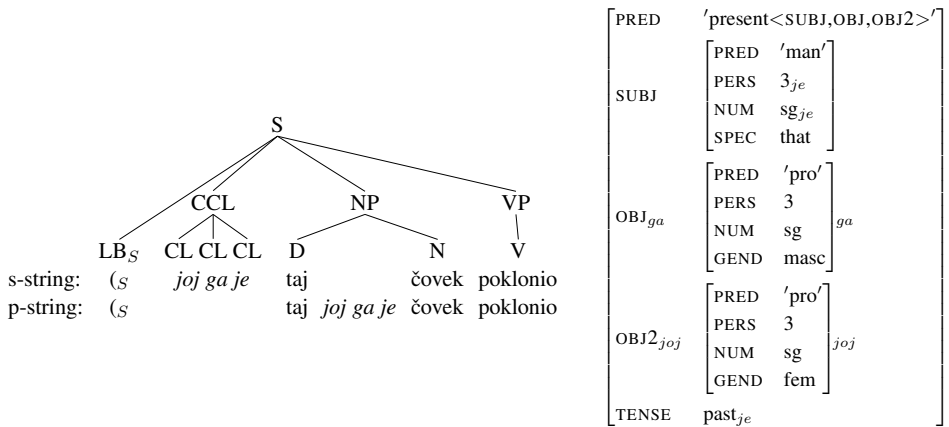
$$(18) \quad \begin{array}{l} \text{s(yntactic)-string ('upper')}: \quad ({}_S \quad CS \quad \omega \quad 0 \\ \hline \text{p(rosodic)-string ('lower')}: \quad ({}_S \quad 0 \quad \omega \quad CS \end{array}$$

The regular relation has the effect that strings with syntactically clause-initial clitic sequences are related to strings where those clusters appear on the other side of an adjacent prosodic word. The sentence-initial position allows for the

<sup>17</sup>For further work in LFG, see an account of prosodically determined second position clitics in Vafsi in Bögel et al. (2018).

3 Prosody and its interfaces

functional information to be made available to syntax, but violates the prosodically dependent clitic’s need for a preceding host. The second position in the prosodic string satisfies this prosodic constraint in that the clitic is placed following a valid prosodic host.



Tina Bögel

#### 4.4 A strictly parallel approach: Dalrymple & Mycock (2011), Mycock & Lowe (2013)

Dalrymple, Mycock, and Lowe base their approaches<sup>18</sup> on the assumption that the prosodic and the syntactic component are parallel but separate components, which goes beyond the distinction between direct and indirect approaches to prosody and its interfaces as briefly discussed in Section 2.2. In the parallel approach, syntactic structure has no influence on the formation of prosodic structure (and vice versa).<sup>19</sup> Instead, each structure is built up independently: syntactic structure as traditionally assumed, and prosodic structure based on rhythmic principles, more specifically, the trochaic foot. In contrast, the indirect reference approach assumes that rhythmic structure is only one factor among many which contribute to the formation of prosodic structure.

The approach represents prosodic constituency in a tree-like structure, assuming the constituents proposed by Selkirk (1995) (see Section 2.1). Similar to the proposal made in Bögel et al. (2010), the interface between the syntactic and prosodic components is the interface between a s(yntactic)-string and a p(honological)-string. The ‘linguistic signal’ (Dalrymple et al. (2019: 407), the nature of which is not further defined) is parsed into minimal syntactic units in the s-string, and into minimal prosodic units (i.e., syllables) in the p-string. A (very simplified) representation of the example sentence *Anna was studying at the university* is shown in Figure 13 (see Figure 15 for a complete picture).

Figure 13 displays the syntactic component (s-string and c-structure) in the top part. The bottom part represents p-structure: the p-string and the prosodic tree. The p-string is parsed into syllables (but see below for further specifications) which are grouped into prosodic words. Following Lahiri & Plank (2010), prosodic structure is built based on rhythmic principles, specifically on the trochaic foot (see Section 2.2). The representation omits the foot structure, but the underlying formation algorithm is still visible in the fact that the left edge of each prosodic word is placed with the syllable carrying primary stress in a lexical, (syntactic) word; e.g., u.ni.(VER.si.ty)<sub>ω</sub>. Function words and feet built on secondary stress (e.g., (u.ni)<sub>ft</sub>) seem to generally be phrased with the preceding

<sup>18</sup>The following approach to the interface was developed in a number of works, namely Mycock (2006); Dalrymple & Mycock (2011); Mycock & Lowe (2013) (see also Lowe 2016 and Jones 2016 for further discussions); the version described here is part of the prosody chapter in Dalrymple et al. 2019.

<sup>19</sup>Dalrymple et al. (2019) classify their approach as indirect reference (p. 398). However, all indirect reference approaches include syntactic structure as a main factor for building up prosodic structure (mostly from the perspective of production). This is not the case here.

### 3 Prosody and its interfaces

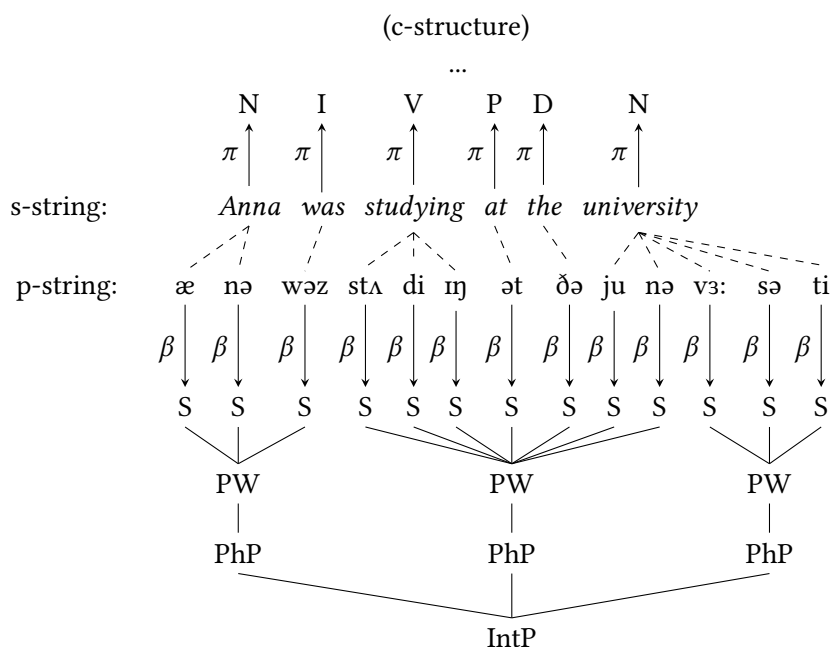


Figure 13: Simplified interface for *Anna was studying at the university*, (Dalrymple et al. 2019: 408, showing only c-structure terminal nodes)

prosodic word.<sup>20</sup> The following example shows the prosodic phrasing according to a trochaic foot structure, with primarily stressed syllables in capital letters.

(19) (Anna was) (STUdying at the uni) (VERsity)

The formation of prosodic words based on rhythmic principles naturally leads to regular mismatches between syntactic and prosodic units. However, the approach raises the question whether these units are indeed prosodic words or whether they should rather be defined as phonological phrases. If these units are prosodic words, then the question arises how phonological phrases are defined under this approach. In Figure 13, each phonological phrase is identical with a prosodic word, which is a crosslinguistically very unusual 1-1 relationship.<sup>21</sup>

<sup>20</sup>It would be interesting to see how this approach can be applied to cases where the first syllable of a prosodic unit is unstressed as in a modified version of example (19), *Anna or Ravi and Karla* ... (Anna  $\vee$  (Ravi  $\wedge$  Karla)), where the prosodic boundary is realised directly after *Anna* (Wagner 2010), while the rhythmic approach would predict for the prosodic boundary to occur after *or*.

<sup>21</sup>See, for example, the family of BINMIN-constraints, which require for a higher prosodic domain to contain more than one unit of a lower prosodic domain (a.o., Ghini 1993; Inkelas & Zec 1995).

Tina Bögel

Mycock & Lowe (2013) extend the string interface by assuming that the string-units are not atomic but should rather be seen as feature bundles, represented as AVMs. The relation between the two strings and their units is regulated through information stored in the lexicon. While the lexical s(yntactic)-form contains the traditional morphosyntactic information, the p(honological)-form contains information on segments and syllable structure as well as the feature SYLLSTRESS which indicates the primary lexical stress position.

In addition to the lexical information, the feature structures at the interface also include information on the edges of constituents in the respective modules. These ‘edge features’ are necessary to allow for the matching of prosodic and syntactic constituents, e.g., in order to prosodically disambiguate syntactically ambiguous structures. Mycock and Lowe define a number of mechanisms to make the edges available to the strings:  $\nearrow$  and  $\searrow$  for the left and right edges of syntactic nodes, and  $\nearrow$  and  $\searrow$  for the left and right edges of prosodic nodes.<sup>22</sup> Figure 14 shows the AVMs for the first syntactic and the first prosodic unit of example (19), where the values of l(ef) and r(ight) consist of a set of syntactic and prosodic nodes whose edges are represented by this particular form.

s-string unit <i>Anna</i>	p-string unit <i>æ</i>
$\left[ \begin{array}{ll} \text{FM} & \text{ANNA} \\ \text{L} & \{\text{IP}, \text{NP}, \text{N}\} \\ \text{R} & \{\text{NP}, \text{N}\} \end{array} \right]$	$\left[ \begin{array}{ll} \text{FM} & \text{æ} \\ \text{SYLLSTRESS} & \text{P} \\ \text{L} & \{\text{INTP}, \text{PHP}, \text{PW}\} \\ \text{R} & \{\} \end{array} \right]$

Figure 14: Feature structure for first unit of the p-string and the s-string (Dalrymple et al. 2019: 412)

At the interface, a ‘Principle of Interface Harmony’ ensures that the best-matching parses between the p-string and the s-string are preferred. Note, however, that the approach does not explain how this preference is implemented.<sup>23</sup>

Furthermore, the question of which syntactic and prosodic constituent edges should be matched, that is, which prosodic boundary type is important to syntax and vice versa, is left for future research (Dalrymple et al. 2019: 419). This is surprising, given the extensive existing literature on the topic, but in a sense it is a necessary consequence of assuming strictly parallel modules as has been discussed in Section 2.2.

<sup>22</sup>See Dalrymple et al. (2019) for the exact definitions.

<sup>23</sup>Lowe (2016) presents a possible implementation of the Principle of Interface Harmony using additional formal power in form of OT constraints (see also Lowe & Belyaev 2015). A critical discussion of this approach can be found in Bögel (2015: Ch.6).

### 3 Prosody and its interfaces

Apart from the interface to syntax, the string interface also serves as an interface between prosody and semantics and i-structure.<sup>24</sup> The semantics-prosody interface is demonstrated by means of declaratives, where the intonational contour distinguishes between declarative statements and questions. In order to make the semantic information available at the string interface, the c-structure receives a ‘label’ *PolarIntSem* along with the meaning constructor [*PolarInt*]. Similar to the edge values, this label is handed down to the s-string where it is placed in the rightmost AVM.

For the prosodic interpretation of a declarative question, information on pitch is required. This information is included in the form of H and L pitch accents and boundary tones.

Dalrymple et al. (2019) assume that in English declarative questions, a nuclear L tone is associated with the stressed syllable of the first prosodic word in the last phonological phrase, and a H boundary tone at the right edge of the Intonational Phrase.<sup>25</sup> Similar to O’Connor (2005) above (Section 4.2), they annotate prosodic structure by means of prosodic phrase structure rules. In addition, a label *PolarInt* appears at the rightmost AVM of the p-string. The rule in (20) can then be read as follows: In this phonological phrase, assign a nuclear tone L to the leftmost syllable with primary stress ( $\sigma^s$ ) and a right boundary tone to the rightmost unit;<sup>26</sup> if these annotations come to pass, create a label *PolarInt* which appears as a set member of the rightmost unit’s right edge.

$$(20) \quad \text{IntP} \longrightarrow \text{PhP}^* \quad \text{PhP} \\
\begin{aligned}
& ((\sigma^s \text{ N\_TONE}) = \text{L}) \\
& (\text{RB\_TONE}) = \text{H} \Rightarrow \\
& \text{PolarInt} \in (\text{R})
\end{aligned}$$

Figure 15 shows part of the full analysis for example (19), the PP *at the university*. In addition to the edge features, both labels, *PolarIntSem* and *PolarInt* appear at the right edge of the string interface. The Principle of Interface Harmony requires both labels to co-occur for the overall structure to be grammatical, but the matching process is not further detailed here.

<sup>24</sup> As the description of the interface to i-structure is similar to the one provided for the prosody-semantics interface, the interested reader is referred to Dalrymple et al. (2019) for details.

<sup>25</sup> The prosodic expression of declarative questions in English shows much more variability than assumed here, see, e.g., Gunlogson (2003) for a discussion of different contours.

<sup>26</sup> The assignment of a right boundary tone here does not distinguish between ‘boundary tones’, e.g. H%, which appear as boundary tones of intonational phrases, and ‘phrase accents’, e.g., H<sup>-</sup>, which appear at the edges of phonological phrases. If both edges fall together, these tones form combinations, e.g. H-H%, which can be crucial for an interpretation. The use of a boundary tone with a phonological phrase unfortunately collapses this distinction.

Tina Bögel

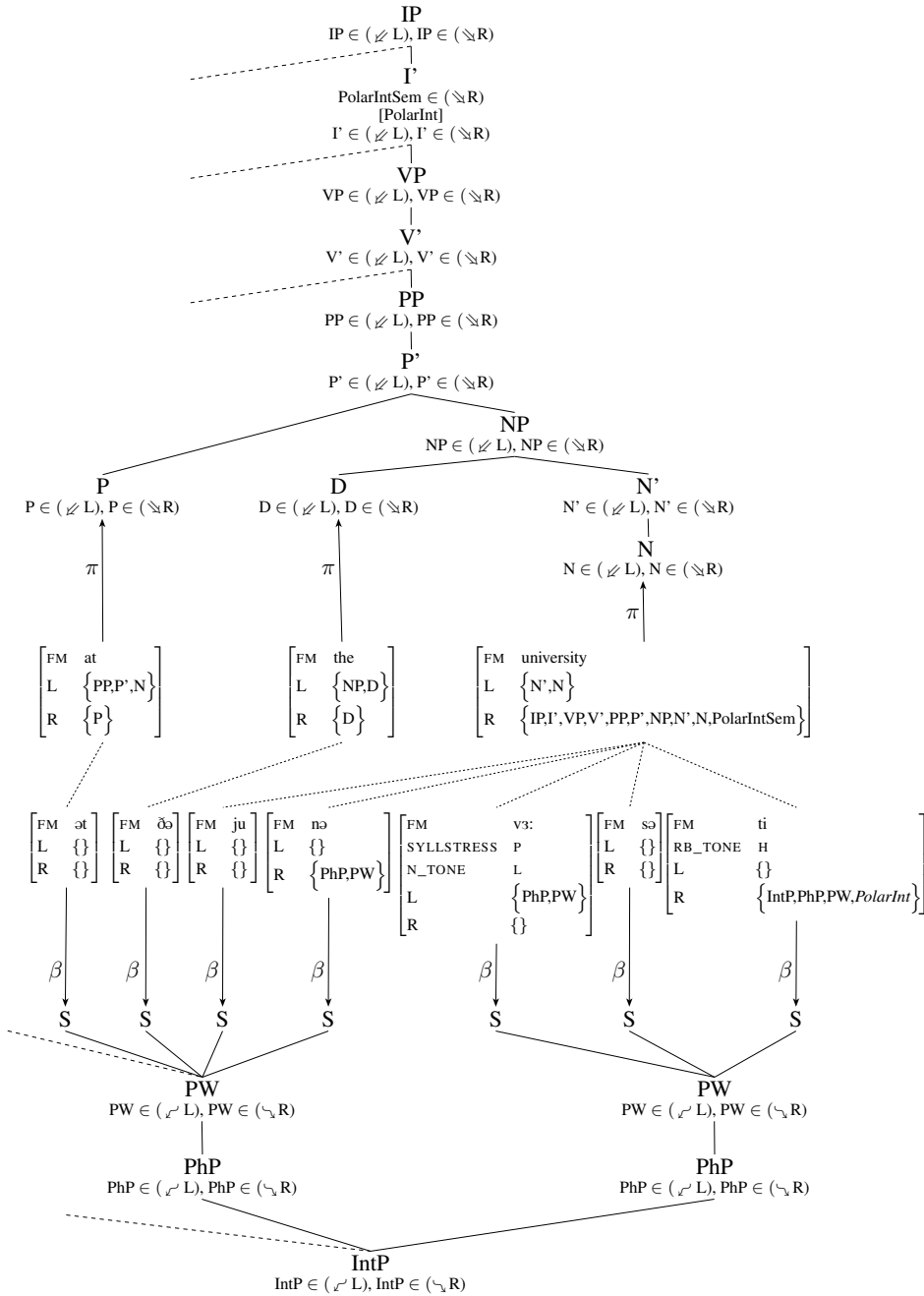


Figure 15: Analysis of the PP in the declarative question *Anna was studying at the university*



### 3 Prosody and its interfaces

In conclusion, the non-directional model proposed by Dalrymple et al. (2019) pursues the idea of modularity in the extreme: the syntactic and the prosodic component are taken to be completely independent structures which do not allow for any co-description mechanisms as commonly found in LFG (see Belyaev 2021a [this volume]). Instead, information about (at least) syntactic and prosodic constituency, semantics, i-structure, and intonational contours is handed to the respective strings. The interface between the syntactic and the prosodic component is then situated between the s-string and the p-string, where matching edges and ‘labels’ are preferred according to the Principle of Interface Harmony.

Apart from initial suggestions involving OT-constraints in Lowe (2016) and Lowe & Belyaev (2015), the formal implementation of the Principle of Interface Harmony is not further detailed. Given that there are numerous combinatorial possibilities of prosodic constituents, pitch accents, phrase accents, and boundary tones, and hardly any of them can be mapped to one particular interpretation, but are always co-dependent on other modules of the grammar, the matching of labels at the string interface will most likely prove to be difficult. The introduction of these labels and the mingling of information from different modules is, however, a necessary consequence of the parallel approach. The reduction of the interface to the strings implies that *all* potentially relevant information from other modules has to be duplicated and appear as part of the string where it might or might not be matched against the material in the parallel string. As it was the case in Bögel et al. (2009, 2010) (Section 4.3), this is also problematic with respect to modularity.

The extensive duplication and blending of structures can be avoided by assuming a more traditional co-descriptive approach, while at the same time acknowledging modularity in that each module only processes information related (i.e., ‘native’) to its module. This indirect reference approach was first pursued in Butt & King (1998), and was further developed in Bögel (2015) and subsequent work as discussed in the next section.

#### 4.5 Production and comprehension: Bögel (2015)

Starting with her dissertation in 2015, Bögel developed a directional indirect reference model of the prosody-syntax interface that enables the integration of a speech signal into LFG and can account for a vast variety of phenomena from both perspectives, production and comprehension. In this approach, the interface between c-structure and p-structure is regulated via two transfer processes, the ‘transfer of vocabulary’ ( $\rho$ ), which exchanges phonological and morphosyntactic information of lexical elements via the multidimensional lexicon, and the

Tina Bögel

‘transfer of structure’ ( $\sharp$ ), which exchanges information on syntactic and prosodic phrasing, and on intonation.

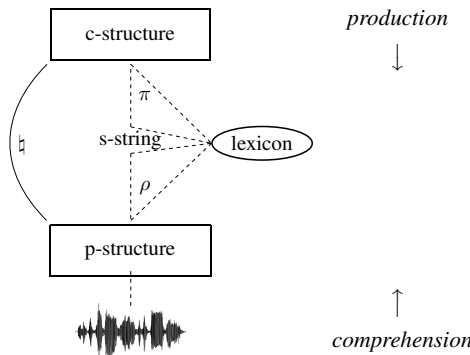


Figure 16: Abstract overview of the prosody-syntax interface (Bögel 2015)

The transfer of vocabulary requires a lexical entry to contain detailed information on (at least) the morphosyntactic as well as the phonological form (s-form and p-form, Dalrymple & Mycock 2011). Following Levelt et al. (1999), Bögel (2015) develops a multidimensional lexicon where the s-form encodes the traditional morphosyntactic information, and the p-form contains information on segments, the metrical frame and prosodic word status.<sup>27</sup> Figure 17 shows the (shortened) lexical entries for *čovek* (‘man’), *taj* (‘that’), and *joj* (‘her’) from example (15) (repeated in (21)) from SCB, where the clitic cluster (*joj ga je*) is placed in the prosodic second position, syntactically ‘interrupting’ the NP *taj čovek*.

- (21) [Taj *joj ga je* čovek] poklonio.  
 that her it AUX man presented  
 ‘That man presented her with it.’ (Schütze 1994)

The lexical p-form entries of *čovek* and *taj* are both marked as full prosodic words ( $\omega$ ). In contrast, the p-form of *joj* is marked as a prosodically deficient enclitic ( $=\sigma$ ), that is, it is prosodically dependent on a preceding host. Following the concept of modularity, each dimension can only be accessed by the related module: c-structure can access the s-form and p-structure the p-form. However, during the transfer of vocabulary, the lexicon also assumes a ‘transducer function’ between s-form and p-form: If a particular dimension is accessed (e.g., s-form from

<sup>27</sup> A third dimension, ‘concept’, which includes semantic information is assumed as well, but not discussed further here.

### 3 Prosody and its interfaces

s(yntactic)-form				p(honological)-form	
čovek	N	(↑ PRED) (↑ PERS) ...	= 'čovek' = 3	P-FORM	[tʃovek]
				SEGMENTS	/tʃ o v e k/
				METR. FRAME	(σσ) <sub>ω</sub>
taj	PRON	(↑ PRED) (↑ PRON-TYPE) ...	= 'pro' = demon	P-FORM	[taj]
				SEGMENTS	/t a j/
				METR. FRAME	(σ) <sub>ω</sub>
joj	PRON	(↑ PRED) (↑ PRON-TYPE) ...	= 'pro' = pers	P-FORM	[joj]
				SEGMENTS	/j o j/
				METR. FRAME	=σ

Figure 17: Lexical entries for SCB *čovek* 'man', *taj* 'that', and *joj* 'her'

c-structure), the associated dimensions become available as well and the information stored in them is projected to their respective structures (e.g., p-form information becomes available to p-structure).

P-structure itself is represented by the p-diagram, a compact linear representation of the utterance. The p-diagram is structured syllablewise, where each syllable is part of a vector (v(ECTOR) INDEX) which associates the syllable with relevant segmental and suprasegmental phonological information.<sup>28</sup> During the transfer of vocabulary, the information stored with each lexical item's p-form is stored in the p-diagram. Figure 18 illustrates this process for example (21).

↑ PHRASING	↑	=σ	=σ	=σ	(σ) <sub>ω</sub>	(σ	σ) <sub>ω</sub>	...	↑
LEX.STRESS		–	–	–	prim	prim	–	...	
SEGMENTS		/joj/	/ga/	/je/	/taj/	/tʃo/	/vek/	...	
V. INDEX		S <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>	S <sub>4</sub>	S <sub>5</sub>	S <sub>6</sub>	...	

Figure 18: The p-diagram with material from the transfer of vocabulary from example (21) (production)

The p-diagram's content in Figure 18 is identical with the lexical p-form information in Figure 17: *tʃovek*, for example, consists of two syllables, each of which contains a number of segments. The first syllable has primary stress and the complete word forms a prosodic word. These attributes and their values are stored

<sup>28</sup>The choice of the underlying prosodic or phonological unit and of the different attributes is up to the researcher.

Tina Bögel

for each syllable, thus creating a linear representation of the phonological string, with a vertical representation of the different values associated with each part of the phonological string.

In addition to the lexical information, p-structure receives information on syntactic constituency through the transfer of structure. This approach assumes match theory<sup>29</sup> (see Section 2.2), where each syntactic clause is mapped to an intonational phrase and each XP is mapped to a phonological phrase. The following c-structure annotation models this approach to the mapping between syntactic and prosodic constituents for the clausal node S.

$$(22) \quad \begin{array}{c} S \\ (* (T(*)) S_{min} \text{ PHRASING}) = {}_l ( \\ (* (T(*)) S_{max} \text{ PHRASING}) = )_l \end{array}$$

This annotation can be read as follows: Take all terminal nodes (T) of the current node (\*, here S), for the attribute PHRASING assign a left IntP boundary ( ${}_l()$ ) to the leftmost syllable ( $S_{min}$ ) and a right IntP boundary to the rightmost syllable ( $S_{max}$ ) in p-structure. The transfer of structure thus encodes information on larger prosodic domains in p-structure. Taken together, the transfer of vocabulary and the transfer of structure thus provide an initial input to p-structure based on lexical phonological information on the one hand, and on syntactic constituency in form of larger prosodic domains on the other hand (see Figure 19).

As discussed above in Section 4.3, the syntactic analysis of example (21) positions the clitic cluster in the sentence-initial position:  $[joj\ ga\ je]_{CCL} [taj\ \check{c}ovek]_{NP} [poklonio]_{VP}$ . As information is accumulated, a prosodic constraint violation becomes apparent (which, in line with modularity, syntax neither recognized nor cared about): The clitics are placed in the initial position of the intonational phrase  $\iota$ , where they cannot attach to a preceding prosodic host:  $(\iota = \sigma$ . This issue is resolved by positing that p-structure is organised according to its own principles and constraints. One of these constraints is prosodic inversion (Halpern 1995), which allows for the clitics to be placed after the first valid prosodic host – in this case  $(/taj/)_{\omega}$ . As a consequence, the initial linear order of the phonological string as depicted in Figure 19 will be adjusted to satisfy the prosodic constraints. The result is the (final) p-string  $(taj)_{\omega} = joj = ga = je\ (t\check{c}ovek)_{\omega} \dots$  which in turn forms the basis for the phonetic representation.

<sup>29</sup>Which approach is chosen for the mapping between syntactic and prosodic constituency is up to the researcher. In this case, the end-based approach would not lead to a different outcome with respect to the clitic placement.

3 Prosody and its interfaces

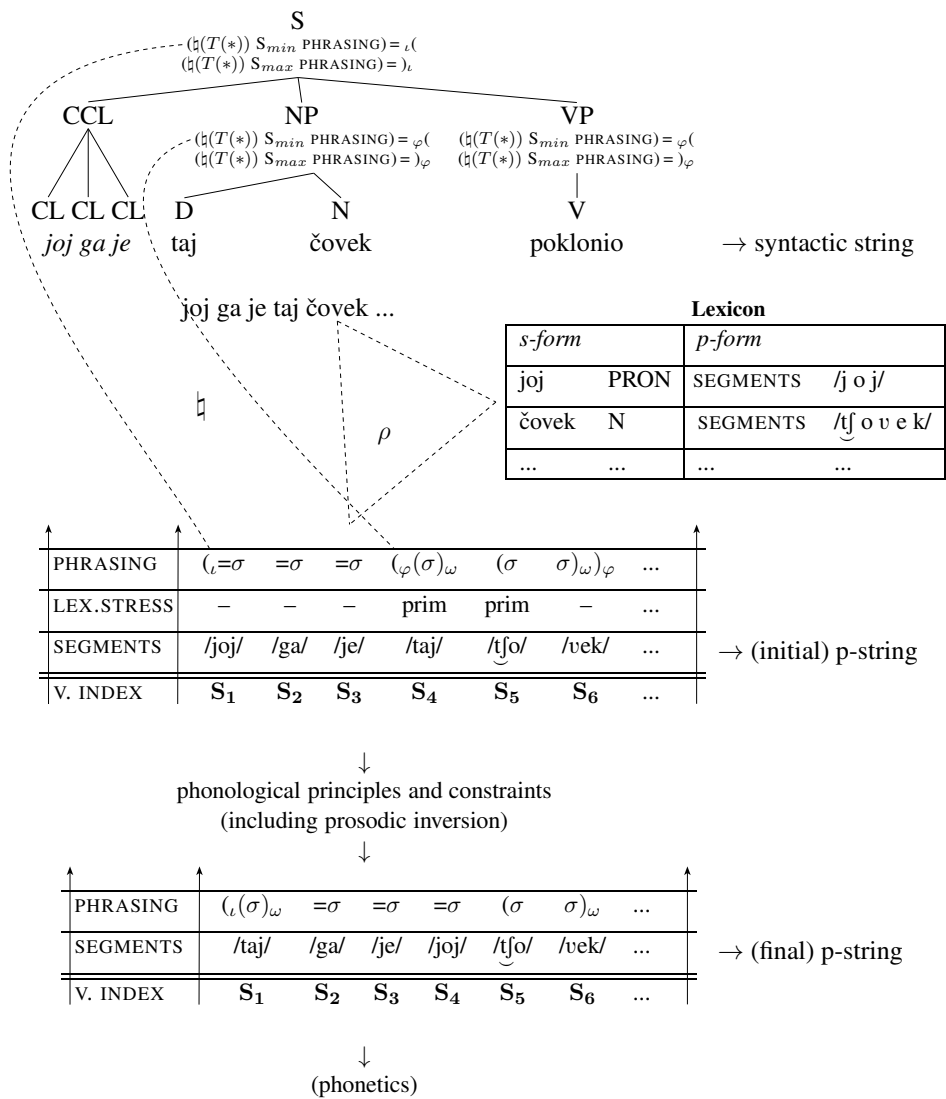


Figure 19: The syntax–prosody interface and the p-structure analysis for example (21)

Tina Bögel

This approach to the interface allows for a clear separation of syntactic and prosodic analyses and can account for a number of other phenomena as well, including notorious problems where the concept of lexical integrity (Bresnan & Mchombo 1995) is seemingly at stake. Such phenomena, among them Pashto endoclititics (Bögel 2015: Ch. 6) and Vafsi mesoclititics (Bögel et al. 2018), are difficult to explain from a purely syntactic perspective, but can be explained in a straightforward fashion with the proposal made here, as prosodic restructuring is based on prosodic constraints alone and is not concerned with syntactic constraints like word integrity.

The 2P clitic analysis just discussed was an example for production, i.e., the analysis first considers syntactic structure and then builds prosodic structure. The framework proposed in Bögel (2015) also allows for comprehension as is demonstrated in the following with an example from Butt et al. (2017, 2020) on Urdu polar *kya*. Consider example (23), where the sentence can be understood either as a polar question or as a wh-constituent question.

- (23) alina=ne zain=ko kya tohfa di-ya t<sup>h</sup>-a?  
Alina=Erg Zain=Acc what present.M.Sg give-Perf.M.Sg be.Past-M.Sg  
Constituent Question: ‘What gift did Alina give to Zain?’  
Polar Question: ‘Did Alina give a gift to Zain?’

This ambiguity corresponds to two different possible syntactic analyses. In the wh-constituent interpretation, *kya* is phrased together with the following noun *tohfa*. In contrast, in the polar interpretation, *kya* is analyzed as an immediate daughter of S.

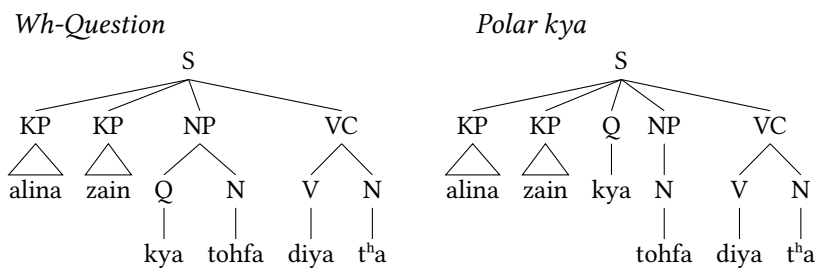


Figure 20: C-structures for the *wh*-reading and for *polar kya*.

Prosody is essential to the disambiguation of this structure: *kya* carries a H\* accent if it is part of a constituent question while it has a flat pitch in the polar interpretation. In order for the grammar to make use of this disambiguation possibility, the information on pitch accents thus has to be available at the interface to p-structure.

### 3 Prosody and its interfaces

The categorical interpretation of pitch accents is dependent on a number of attributes in a given speech signal; for a H\* pitch accent, there needs to be a sudden rise followed by a relatively abrupt drop in the fundamental frequency  $F_0$ . The following p-diagram allows for the integration of this (and additional) speech signal information on the ‘signal’ level (here: medium  $F_0$  and duration for each syllable) with a categorical interpretation of the relevant acoustic cues given on the ‘interpretation’ level in form of a ToBI annotation: H\*.<sup>30</sup>

↑	↑												↑	
PHRAS.	(	...	...	...	...	...	...	...	...	...	...	...	)	INTERPRET.
ToBI	...	...	...	...	...	...	H*	...	...	...	...	L%		↓
DUR.	0,08	0,16	0,14	0,17	0,28	0,23	0,21	0,20	0,16	0,13	0,11	0,22		SIGNAL
$F_0$	164	211	239	243	228	229	247	229	162	147	136	(83)		↓
VALUE	[ə]	[li]	[na]	[ne]	[zæn]	[ko]	[kja]	[təh]	[fa]	[dɪ]	[ja]	[tʰa]		
INDEX	S <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>	S <sub>4</sub>	S <sub>5</sub>	S <sub>6</sub>	S <sub>7</sub>	S <sub>8</sub>	S <sub>9</sub>	S <sub>10</sub>	S <sub>11</sub>	S <sub>12</sub>		

Figure 21: The p-diagram for the speech signal corresponding to the constituent question in (23).

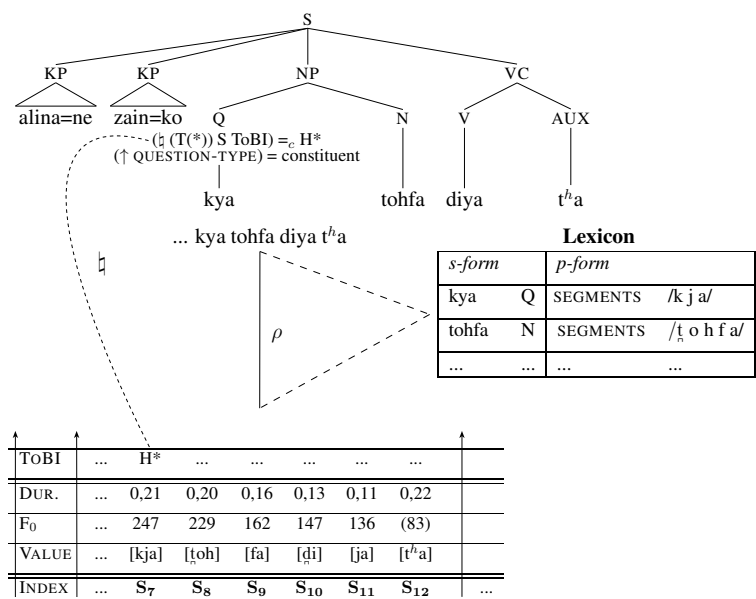
This effectively provides for an interface between phonetics (i.e., a concrete speech signal) and phonology/prosody and allows for the integration of information given in a speech signal into an LFG grammar. The relevant information on the H\* accent again becomes available at the interface to c-structure via the transfer of structure and the transfer of vocabulary.

As shown in Figure 22, the transfer of vocabulary identifies the correct p-forms in the multidimensional lexicon and makes the associated s-forms available to c-structure. The transfer of structure then ‘checks’ whether the syllable associated with the c-structure node Q carries an H\* accent – in which case the attribute-value pair [QUESTION-TYPE = constituent] is projected to f-structure. This approach also allows for the disambiguation of syntactically ambiguous structures where the boundary domains are the crucial indicators (Bögel 2020) and has also been applied to more complex pitch accent phenomena (Bögel & Raach 2020).

Summing up, the directional indirect reference approach proposed in Bögel (2015) allows for a differentiation of production and comprehension and has been

<sup>30</sup>The annotation on the interpretation level is greatly reduced on purpose. There is not yet a fully developed ‘UrduToBI’ or a clear conception of possible prosodic domains and how these are defined (but see Urooj et al. 2019 for a discussion) – an interpretation in terms of the English/German annotation system might thus be misleading.

Tina Bögel



applied to a large variety of different linguistic phenomena. It is the first approach in LFG which integrates spoken language in the form of concrete speech signal data and which pushes LFG towards a more psycholinguistic model of language as discussed in Section 3.

## 5 Conclusion

This chapter gave a chronological overview of the different approaches to prosody and its interfaces in LFG. As the different proposals show, work at this particular interface always requires a discussion of grammar architecture in general and of module interaction in particular. In general, two schools of thought can be distinguished in the LFG literature: the indirect reference approach and the parallel approach. The indirect reference approach assumes that p-structure is influenced by information from different modules, for example syntactic constituency. In addition, p-structure is assumed to be subject to its own principles and constraints, among them rhythmic principles, prosodic inversion, or constraints on the size of prosodic domains. The indirect reference approach was pursued in Butt & King (1998), O'Connor (2005), and Bögel (2015) and subsequent papers. While



### 3 Prosody and its interfaces

these proposals show differences with respect to the overall architecture, the interfaces between modules in all of these approaches are organised according to the traditional co-descriptive LFG annotations.

The second school of thought, the parallel approach, assumes that modules are built up in parallel. Under this view, each module is built on its own principles and constraints without ‘input’ from the other modules. P-structure is assumed to be formed on rhythmic principles, thus accounting for the mismatches found between prosodic and syntactic constituency. The interface between c- and p-structure is reduced to the interface between the syntactic and the phonological string, which are extended to include prosodic and syntactic/semantic information. The information present in both strings is then ‘matched’. This approach is most prominently pursued in Dalrymple & Mycock (2011), Mycock & Lowe (2013), and subsequent work. Bögel et al. (2009, 2010) also fall into this second group. However, the exact nature of p-structure is never defined under this approach and it is thus harder to demarcate.

A second main point of this chapter was that the majority of the proposals presented here assume a certain directionality, which is also in line with psycholinguistic and computational approaches (Forst & King 2021 [this volume]): ‘production’ in the case of Butt & King (1998), ‘comprehension’ in Bögel et al. (2009, 2010); O’Connor (2005), and an open discussion of both in Bögel (2015). This distinction is not evident in the proposal made by Dalrymple & Mycock (2011) and Mycock & Lowe (2013), which represent a perspective where each module builds structure independently of the other modules. The discrepancy between these views of the grammar architecture and of the analysis of language in general has, to my knowledge, not yet been openly debated. This chapter hopefully contributes to a more general discussion of grammar architectures in that it aims to show in a very concrete way what each school of thought pursues and how these ideas can be realised.

## Acknowledgements

I would like to thank Tracy H. King, Miriam Butt, and two anonymous reviewers for their detailed comments, and Mary Dalrymple for her infinite patience.

## References

Asudeh, Ash. 2009. Reflexives in the correspondence architecture. Slides of a talk presented at the University of Iceland.

Tina Bögel

- Baumann, Stefan, Johannes Becker, Martine Grice & Doris Mücke. 2007. Tonal and articulatory marking of focus in German. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*.
- Beckman, Mary E., Julia Hirschberg & Stefanie Shattuck-Hufnagel. 2005. The original ToBI system and the evolution of the ToBI framework. In Sun-Ah Jun (ed.), *Prosodic typology: The phonology of intonation and phrasing*, chap. 2, 9–54. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199249633.003.0002.
- Beckman, Mary E. & Janet B. Pierrehumbert. 1986. Intonational structure in English and Japanese. *Phonology Yearbook* 3. 255–309.
- Belyaev, Oleg. 2021a. Core concepts of LFG. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 21–92. Berlin: Language Science Press. DOI: ??.
- Belyaev, Oleg. 2021b. Introduction to LFG. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 3–20. Berlin: Language Science Press. DOI: ??.
- Bennett, Ryan & Emily Elfner. 2019. The syntax-prosody interface. *Annual Review of Linguistics* 5. 151–171. DOI: 10.1146/annurev-linguistics-011718-012503.
- Bögel, Tina. 2015. *The syntax-prosody interface in Lexical Functional Grammar*. Konstanz: University of Konstanz. (Doctoral dissertation).
- Bögel, Tina. 2020. German case ambiguities at the interface: Production and comprehension. In Gerrit Kentner & Joost Kremers (eds.), *Prosody in syntactic encoding* (Linguistische Arbeiten 573), 51–84. Berlin: De Gruyter.
- 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, CA: CSLI Publications.
- Bögel, Tina, Miriam Butt, Ronald M. Kaplan, Tracy Holloway King & John T. Maxwell III. 2010. Second position and the prosody-syntax interface. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 106–126. Stanford, CA: CSLI Publications.
- Bögel, Tina, Miriam Butt & Sebastian Sulger. 2008. Urdu *Ezafe* and the morphology-syntax interface. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 129–149. Stanford, CA: CSLI Publications.
- Bögel, Tina & Lea Raach. 2020. Swabian *ed* and *edda*: Negation at the interfaces. In Miriam Butt & Ida Toivonen (eds.), *Proceedings of the LFG '20 conference*. Stanford, CA: CSLI Publications.
- Bögel, Tina, Saeed Reza Yousefi & Mahinnaz Mirdehghan. 2018. Vafsi oblique pronouns: Stress-related placement patterns. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '18 conference*, 88–108. Stanford, CA: CSLI Publications.

### 3 Prosody and its interfaces

- Bresnan, Joan & Sam A. Mchombo. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language & Linguistic Theory* 13(2). 181–254. DOI: 10.1007/bf00992782.
- Butt, Miriam, Tina Bögel & Farhat Jabeen. 2017. Polar *kya* and the prosody-syntax-pragmatics interface. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '17 conference*. Stanford, CA: CSLI Publications.
- Butt, Miriam, Farhat Jabeen & Tina Bögel. 2020. Ambiguity resolution via the syntax-prosody interface: The case of *kya* ‘what’ in Urdu/Hindi. In Gerrit Kentner & Joost Kremers (eds.), *Prosody in syntactic encoding*, vol. 573 (Linguistische Arbeiten), 85–118. Berlin: De Gruyter.
- Butt, Miriam & Tracy Holloway King. 1998. Interfacing phonology with LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '98 conference*. Stanford, CA: CSLI Publications.
- Chen, Matthew Y. 1987. The syntax of Xiamen tone sandhi. *Phonology Yearbook* 4. 109–149. DOI: 10.1017/s0952675700000798.
- Dalrymple, Mary. 2001. *Lexical Functional Grammar*. Vol. 34 (Syntax and Semantics). New York: Academic Press. DOI: 10.1163/9781849500104.
- Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III & Annie Zaenen (eds.). 1995. *Formal issues in Lexical-Functional Grammar*. Stanford, CA: CSLI Publications.
- Dalrymple, Mary, John J. Lowe & Louise Mycock. 2019. *The Oxford reference guide to Lexical Functional Grammar*. Oxford: Oxford University Press. DOI: 10.1093/oso/9780198733300.001.0001.
- Dalrymple, Mary & Louise Mycock. 2011. The prosody-semantics interface. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '11 conference*, 173–193. Stanford, CA: CSLI Publications.
- Elordieta, Gorka. 2008. An overview of theories of the syntax-phonology interface. *Anuario del Seminario de Filología Vasca “Julio de Urquijo”* 42(1). 209–286.
- Féry, Caroline. 2020. Grammatical reflexes of information structure in Germanic languages. In Michael T. Putnam & B. Richard Page (eds.), *Cambridge handbook for Germanic linguistics*, 661–685. Cambridge, UK: Cambridge University Press. DOI: 10.1017/9781108378291.029.
- Forst, Martin & Tracy Holloway King. 2021. Computational implementations and applications. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 141–180. Berlin: Language Science Press. DOI: ??.
- Frota, Sónia. 2012. Prosodic structure, constituents, and their implementation. In Abigail C. Cohn, Cécile Fougeron & Marie K. Huffman (eds.), *The Oxford*

Tina Bögel

- handbook of laboratory phonology*, chap. 11, 255–265. Oxford: Oxford University Press.
- Ghini, Mirco. 1993. Phi-formation in Italian: a new proposal. *Toronto Working Papers in Linguistics* 12(2). 41–78.
- Goldsmith, John. 1976. *Autosegmental phonology*. Cambridge, MA: Massachusetts Institute of Technology. (Doctoral dissertation).
- Grice, Martine & Stefan Baumann. 2002. Deutsche Intonation und GToBI. *Linguistische Berichte* 191. 267–298.
- Grice, Martine, D. Robert Ladd & Amalia Arvaniti. 2000. On the place of phrase accents in intonational phonology. *Phonology* 17(2). 143–185. DOI: 10.1017/S0952675700003924.
- Gunlogson, Christine. 2003. *True to form: Rising and falling declaratives as questions in English*. New York: Routledge. DOI: 10.4324/9780203502013.
- Halpern, Aaron. 1995. *On the placement and morphology of clitics*. Stanford, CA: CSLI Publications.
- Hayes, Bruce. 1989. The prosodic hierarchy in meter. In Paul Kiparsky & Gilbert Youmanns (eds.), *Rhythm and meter*, 201–260. Orlando, FL: Academic Press. DOI: 10.1016/b978-0-12-409340-9.50013-9.
- Hayes, Bruce & Aditi Lahiri. 1991. Bengali intonational phonology. *Natural Language & Linguistic Theory* 9(1). 47–96. DOI: 10.1007/bf00133326.
- Inkelas, Sharon & Draga Zec. 1995. Syntax-phonology interface. In John A. Goldsmith (ed.), *The handbook of phonological theory*, chap. 15, 535–549. Cambridge, MA: Blackwell.
- Jackendoff, Ray. 2002. *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.
- Jackendoff, Ray. 2010. The Parallel Architecture and its place in cognitive science. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, chap. 23, 583–605. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199544004.013.0023.
- Jones, Stephen. 2016. The syntax–prosody interface in Korean: Resolving ambiguity in questions. In Doug Arnold, Miriam Butt, Berthold Cysmann, Tracy Holloway King & Stefan Müller (eds.), *Proceedings of the joint 2016 conference on Head-Driven Phrase Structure Grammar and Lexical Functional Grammar*. Stanford, CA: CSLI Publications.
- Kaisse, Ellen M. 1985. *Connected speech: The interaction of syntax and phonology*. Orlando: Academic Press.
- Kaplan, Ronald M. 1987. Three seductions of computational psycholinguistics. In Peter Whitelock, Mary McGee Wood, Harold L. Somers, Rod Johnson & Paul Bennett (eds.), *Linguistic theory and computer applications*, 149–188. London:

## 3 Prosody and its interfaces

- Academic Press. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 339–367).
- Kaplan, Ronald M. & Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics* 20. 331–478.
- Kingston, John. 2019. The interface between phonetics and phonology. In William F. Katz & Peter F. Assmann (eds.), *The Routledge handbook of phonetics*, 359–400. Abingdon & New York: Routledge. DOI: 10.4324/9780429056253-14.
- Kügler, Frank, Stefan Baumann, Bistra Andreeva, Bettina Braun, Martine Grice, Jana Neitsch, Oliver Niebuhr, Jörg Peters, Christine T. Röhr, Antje Schweitzer & Petra Wagner. 2019. Annotation of German intonation: DIMA compared with other annotation systems. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (eds.), *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*. Australasian Speech Science & Technology Association Inc.
- Ladd, D. Robert. 1986. Intonational phrasing: The case for recursive prosodic structure. *Phonology Yearbook* 3. 311–340. DOI: 10.1017/s0952675700000671.
- Lahiri, Aditi & Jennifer Fitzpatrick-Cole. 1999. Emphatic clitics and focus intonation in Bengali. In René Kager & Wim Zonneveld (eds.), *Phrasal phonology*, 119–144. University of Nijmegen Press.
- Lahiri, Aditi & Frans Plank. 2010. Phonological phrasing in Germanic: The judgement of history, confirmed through experiment. *Transactions of the Philological Society* 108(3). 372–398. DOI: 10.1111/j.1467-968x.2010.01246.x.
- Lehiste, Ilse, Joseph P. Olive & Lynn A. Streeter. 1976. Role of duration in disambiguating syntactically ambiguous sentences. *The Journal of the Acoustical Society of America* 60. 1199–1202. DOI: 10.1121/1.381180.
- Levelt, Willem J. M. 1999. Models of word production. *Trends in Cognitive Sciences* 3(6). 223–232. DOI: 10.1016/s1364-6613(99)01319-4.
- Levelt, Willem J. M., Ardi Roelofs & Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22. 1–75. DOI: 10.1017/s0140525x99001776.
- Liberman, Mark. 1975. *The intonational system of English*. Cambridge, MA: Massachusetts Institute of Technology. (Doctoral dissertation).
- Lowe, John J. 2016. Clitics: Separating syntax and prosody. *Journal of Linguistics* 52. 375–419.
- Lowe, John J. & Oleg Belyaev. 2015. Clitic positioning in Ossetic. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '15 conference*. Stanford, CA: CSLI Publications.
- McCarthy, John J. & Alan Prince. 1993. Generalized alignment. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology*, 79–153. Dordrecht: Kluwer Academic Publishers. DOI: 10.1007/978-94-017-3712-8\_4.

Tina Bögel

- McCawley, James D. 1968. *The phonological component of a grammar of Japanese*. The Hague: Mouton.
- Mycock, Louise. 2006. *The typology of wh-questions*. Manchester: University of Manchester. (Doctoral dissertation).
- Mycock, Louise & John J. Lowe. 2013. The prosodic marking of discourse functions. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '13 conference*, 440–460. Stanford, CA: CSLI Publications.
- Nespor, Marina & Irene Vogel. 1986. *Prosodic phonology*. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110977790.
- Nespor, Marina & Irene Vogel. 1989. On clashes and lapses. *Phonology* 6(1). 69–116. DOI: 10.1017/s0952675700000956.
- O'Connor, Robert. 2005. Clitics in LFG: Prosodic structure and phrasal affixation. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '05 conference*. Stanford, CA: CSLI Publications.
- Pierrehumbert, Janet B. 1980. *The phonology and phonetics of English intonation*. Cambridge, MA: Massachusetts Institute of Technology. (Doctoral dissertation).
- Pierrehumbert, Janet B. & Mary E. Beckman. 1988. *Japanese tone structure*. Cambridge, MA: The MIT Press.
- Prince, Alan & Paul Smolensky. 2004. *Optimality theory: Constraint interaction in generative grammar*. Oxford/Malden: Blackwell.
- Sadock, Jerrold M. 1991. *Autolexical Syntax: A theory of parallel grammatical representations*. Chicago: University of Chicago Press.
- Schütze, Carson. 1994. Serbo-Croatian second position clitic placement and the phonology-syntax interface. In Andrew Carnie, Heidi Harley & T. Bures (eds.), *MIT Working Papers in Linguistics: Papers on phonology and morphology*, 373–473. Cambridge, MA: Department of Linguistics & Philosophy, MIT. Revised version.
- Selkirk, Elisabeth O. 1978. On prosodic structure and its relation to syntactic structure. In Thorstein Fretheim (ed.), *Nordic prosody ii*, 111–140. Tapir.
- Selkirk, Elisabeth O. 1984. *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: The MIT Press.
- Selkirk, Elisabeth O. 1986. On derived domains in sentence phonology. *Phonology Yearbook* 3. 371–405. DOI: 10.1017/s0952675700000695.
- Selkirk, Elisabeth O. 1995. The prosodic structure of function words. In Jill N. Beckmann, Laura W. Dickey & Suzanne Urbanczyk (eds.), *Papers in optimality theory*. University of Massachusetts: Department of Linguistics.

### 3 Prosody and its interfaces

- Selkirk, Elisabeth O. 2009. On clause and intonational phrase in Japanese: The syntactic grounding of prosodic constituent structure. *Gengo Kenkyu* 136. 35–73.
- Selkirk, Elisabeth O. 2011. The syntax-phonology interface. In John A. Goldsmith, Jason Riggle & Alan C. L. Yu (eds.), *The handbook of phonological theory*, 435–484. Malden, MA: Blackwell. DOI: [10.1002/9781444343069.ch14](https://doi.org/10.1002/9781444343069.ch14).
- Shattuck-Hufnagel, Stefanie & Alice Turk. 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research* 25(2). 193–247. DOI: [10.1007/bf01708572](https://doi.org/10.1007/bf01708572).
- Silverman, Kim, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet B. Pierrehumbert & Julia Hirschberg. 1992. ToBI: a standard for labeling English prosody. In *Proceedings of the 1992 international conference on spoken language processing*. Banff.
- Sweet, Henry. 1904. *Elementarbuch des gesprochenen Englisch*. 3rd edn. Leipzig: Tauchnitz [u.a.]
- Urooj, Saba, Benazir Mumtaz & Sarmad Hussain. 2019. Urdu intonation. *Journal of South Asian Linguistics* 10. Special issue on the prosody of South Asian languages.
- Wagner, Michael. 2010. Prosody and recursion in coordinate structures and beyond. *Natural Language & Linguistic Theory* 28. 183–237. DOI: [10.1007/s11049-009-9086-0](https://doi.org/10.1007/s11049-009-9086-0).
- Wells, John C. 1970. Local accents in England and Wales. *Journal of Linguistics* 6(2). 231–252. DOI: [10.1017/s0022226700002632](https://doi.org/10.1017/s0022226700002632).
- Zaenen, Annie & Ronald M. Kaplan. 1995. Formal devices for linguistic generalizations: West Germanic word order in LFG. In Jennifer S. Cole, Georgia M. Green & Jerry L. Morgan (eds.), *Linguistics and computation*, 3–27. Stanford, CA: CSLI Publications. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 215–240).





## **Part III**

# **Formal and computational issues and applications**



## Chapter 4

# Computational implementations and applications

Martin Forst

Cerence Inc.

Tracy Holloway King

Adobe Inc.

Computational implementations of LFG are computer programs composed of LFG annotated c-structure rules and lexical entries. LFG was designed to be computationally tractable and has a strong history of broad-coverage grammar implementations for diverse languages. As with theoretical LFG, implemented grammars primarily focus on c-structure and f-structure, but the resulting f-structures are used as input to semantics and abstract knowledge representation, and some work has focused on the integration of morphological and phonological information as well as argument structure. From a theoretical linguistic perspective, implemented grammars allow the linguist to test analyses and to see interactions between different parts of the grammar. From an application perspective, applications such as machine translation and question answering take advantage of the abstract f-structures and the ability of LFG grammars to parse and generate as well as to detect (un)grammaticality.

Computational implementations of LFG are computer programs composed of LFG annotated c-structure rules and lexical entries. When parsing, they take as input a natural language sentence and output c-structures and f-structures and potentially other projections such as semantics. When generating, they take as input an f-structure and generate a grammatical natural language sentence. As with theoretical LFG, these implemented grammars obey the fundamental premises of LFG such as completeness, coherence, and uniqueness.<sup>1</sup> LFG was de-

---

<sup>1</sup>This contrasts with approaches which produce f-structure-like representations but do not use LFG principles or machinery. See Section 3 and Cahill et al. 2002.



*Martin Forst & Tracy Holloway King*

signed from the outset to be computationally tractable and has a strong history of broad-coverage implementations for multiple languages, primarily through the ParGram project (Butt, King, Niño, et al. 1999) which is built on the XLE grammar development platform (Crouch et al. 2008).

Grammar engineering involves the implementation of linguistically-motivated grammars so that natural language utterances and text can be processed to produce deep syntactic, and sometimes semantic, structures. As with theoretical LFG, implemented grammars primarily focus on c-structure and f-structure. The resulting f-structures have been used extensively as input to semantics and abstract knowledge representation. Other work has focused on the integration of morphological and phonological information, as well as argument structure, but in general these areas have lagged behind the proposals in the theoretical literature. In addition, implemented LFG grammars have been used to create large-scale tree and dependency banks, mapping a corpus of sentences to a set of f-structures or related dependency structures.

We first introduce the computational implementations of LFG, presenting specific platforms and touching upon aspects such as core components, grammar development tools, modularity, and runtime performance (Section 1). We then discuss implications for theoretical issues (Section 2) and the ParGram grammar resources (Section 3). Finally, we outline existing and potential applications for LFG implemented grammars (Section 4).

## 1 Computational Implementations

Computational implementations of LFG grammars focus on annotated phrase structure rules and lexical entries. These implementations concentrate on creating high-quality f-structures since most applications use f-structures as their input (Section 4). This section first introduces the major platforms that support LFG implementations. The core components provided by these platforms are then outlined, followed by some specific grammar development tools. Finally two computational notions, modularity and performance, are discussed.

### 1.1 Platforms

Since the inception of LFG as a grammar framework several platforms aimed at processing text according to the LFG formalism have been created. These platforms range from an M.Sc. project (Minos 2014) and introductory French implementation (Zweigenbaum 1991) to an industrially funded grammar development

## 4 Computational implementations and applications

and processing platform which was actively developed for over two decades: the Xerox Linguistic Environment (XLE). In between those in terms of breadth of applicability and technical maturity are systems developed in academic research institutions, in particular XLFG, SxLFG, and the Free Linguistic Environment (FLE). Active development on many of these systems is limited: for current status and documentation the platform owners should be consulted.

### 1.1.1 XLFG and Elvex

XLFG (Clément & Kinyon 2001) is a parsing platform that was first implemented for didactic purposes.<sup>2</sup> It has been used to verify the soundness of several proposals to handle a variety of linguistic phenomena (Section 2), e.g. zeugmas, particle verbs, and non-constituent coordination (Clément 2019).

XLFG uses an Earley parser (Earley 1970) for context-free parsing, and then resolves the f-structure constraints on packed c-structure representations (Maxwell & Kaplan 1989, 1993). It expects tokenized sentences as input and uses full-form lexicons for lexical lookup (Section 1.2). XLFG does not facilitate the use of external components like finite-state transducers for preprocessing tasks such as tokenization or morphological analysis (Section 1.2). It has primarily been applied to parsing French and English, i.e. analyzing French or English text into f-structures. Recently, work was started on a generator, i.e. mapping f-structures to text, using XLFG-style grammars for the production of surface realizations from f-structures. This generator is named Elvex.<sup>3</sup>

### 1.1.2 SxLFG

SxLFG (Boullier & Sagot 2005) was also developed with the participation of Lionel Clément, but its main authors are Pierre Boullier and Benoît Sagot of INRIA. The primary focus of SxLFG is on the deep non-probabilistic parsing of large corpora (Sagot & Boullier 2006) by means of robustness techniques for input sentences for which no spanning c-structure can be produced. The underlying context-free parser is the Early parser of the SYNTAX project. Like XLFG and XLE, SxLFG resolves f-structure constraints on packed c-structure representations. The French broad-coverage LFG implementation that has been used extensively with SxLFG includes a large full-form lexicon for French, the Lefff 2 (Sagot et al. 2006). Like XLFG, SxLFG does not facilitate the use of external components

---

<sup>2</sup>XLFG is available at <http://www.xlfg.org>

<sup>3</sup>Elvex is available at <https://github.com/lionelclement/Elvex>

*Martin Forst & Tracy Holloway King*

like finite-state transducers for preprocessing tasks such as tokenization or morphological analysis (Section 1.2). SxLFG was developed for parsing. Generation has not been in the scope of SxLFG.

### 1.1.3 XLE (and GWW as precursor) and XLE-Web

The Xerox Linguistic Environment (XLE) was developed by the Natural Language Theory and Technology (NLTT) group at the Xerox Palo Alto Research Center (PARC). It started as a reimplementaion in C of the earlier Grammar Writer’s Workbench (GWW) (Kaplan & Maxwell 1996), which was implemented in Lisp and is still available. XLE was used by several academic and industry teams for the development of LFG implementations for more than a dozen languages (see Section 3 on the ParGram project). XLE, in conjunction with a customized broad-coverage grammar, was used to parse the English Wikipedia in the Powerset search engine (Kaplan 2009).

XLE has mostly been used for parsing, but it includes a generator that can efficiently produce surface realizations from f-structures and even packed f-structure charts (Maxwell 2006). Thanks to this bidirectionality, it has powered applications such as machine translation and sentence condensation (Section 4), and it has been used in research projects on stochastic realization ranking (Cahill & Forst 2009).

From its inception, XLE was designed to use finite-state transducers for low-level processing steps such as (de)tokenization and morphological analysis and generation (Section 1.2). The interface can readily be used with transducers in Xerox’s finite-state transducer format including ones converted from the Foma finite-state transducers, and with relatively little programming effort, other external components can be integrated into an XLE grammar (e.g. see Fang & King 2007 on integrating a non-finite state Chinese word breaker into an XLE Chinese grammar).

In addition to the  $\phi$ -projection from c-structures to f-structures, the XLE parser supports further projections from either of those representations. One of them, the optimality structure, is hard-coded to guide the parsing and generation process on the basis of optimality marks (Section 1.5). The use of optimality marks as a robustness mechanism is one of the many extensions of XLE born out of a joint effort of the group at PARC and its ParGram partners.

Other extensions of the parser and generator are aimed at reducing latency and at ranking the (top n) parses or realizations. For the former, the most notable mechanism is c-structure pruning (Cahill et al. 2008). C-structure pruning relies on corpus data annotated with (partial) constituent bracketing and learns

## 4 Computational implementations and applications

to eliminate highly unlikely c-structures before the computationally expensive resolution of f-annotations. For the latter, a component for training<sup>4</sup> and applying maximum-entropy models based on a large variety of features is provided as part of XLE (Riezler et al. 2002).

Beyond the parser and the generator, XLE also contains a term-rewriting component which was first developed for transfer-based machine translation but has been used for a number of other purposes: identifying and deleting modifiers in f-structures that can be deleted without changing the meaning of the corresponding sentences too much (Riezler et al. 2003); treebank (**chapters/Treebanks**) conversion from one dependency-oriented format into another (Forst 2003); further normalization of f-structures and/or construction of semantic representations (Crouch & King 2006; Bobrow et al. 2007); extraction of features for parse ranking (Forst 2007) and realization ranking (Cahill & Forst 2009).

Currently XLE is used by the academic members of the ParGram initiative (Section 3) as well as by individual researchers. It can be used online with LFG implementations for a number of languages via XLE-Web,<sup>5</sup> a web interface for XLE developed at the University of Bergen, and is used as part of the ParseBanker infrastructure developed there (Rosén et al. 2009, 2012). See **chapters/Treebanks**. XLE is available for non-commercial research purposes.<sup>6</sup> Uses beyond that require a license agreement with PARC and Xerox.

### 1.1.4 FLE

The Free Linguistic Environment (FLE) (Ĉavar et al. 2016) aims to create an LFG-oriented grammar-development and parsing environment with a license less restrictive than XLE's. It is implemented in C++ and uses the same grammar syntax as XLE, but it is subject to the Apache 2.0 license. In addition to the context-free grammar format of XLE, it supports two probabilistic context-free grammar formats. For tokenization and morphological analysis, FLE provides an interface to Foma transducers.<sup>7</sup> FLE uses open-source components when possible. FLE provides basic parsing functionality but does not contain a generator capable of producing surface strings for input f-structures.<sup>8</sup>

---

<sup>4</sup>Training data comprises sentences with labeled bracketing, which can be derived from treebanks or created manually (Riezler et al. 2002).

<sup>5</sup>XLE-Web is available at <http://clarino.uib.no/iness/xle-we>

<sup>6</sup>XLE is available at <https://ling.sprachwiss.uni-konstanz.de/pages/xle/redmine.html>

<sup>7</sup>Foma supports the import from and the export to XFST formats and XFST supports Foma transducers.

<sup>8</sup>FLE is available at <https://gorilla.linguistlist.org/fle/>

*Martin Forst & Tracy Holloway King*

## 1.2 Core components

The LFG systems described above allow grammar writers to implement LFG grammars with annotated phrase structure rules and lexical entries similar to those in theoretical LFG. The main difference is that the formatting is specified with easier-to-type variants, e.g. symbols like  $\uparrow$  and  $\downarrow$  are replaced with  $\wedge$  and  $!$ .

- (1) Example theoretical and implemented annotated c-structure rules:

Theoretical notation:	Implementation (XLE system) notation:
S $\rightarrow$ NP VP	S $\rightarrow$ NP: ( $\wedge$ SUBJ)=!;
( $\uparrow$ SUBJ)= $\downarrow$ $\uparrow$ = $\downarrow$	VP: $\wedge$ =!.

### 1.2.1 Pre-processing

In order to implement an LFG grammar, it is necessary to preprocess the text that the grammar will parse. Minimally the preprocessing contains a tokenizer which breaks the text into tokens (i.e. words) and canonicalizes the capitalization if necessary (e.g. lowercasing sentence initial capitalized words in English unless they are proper nouns). These canonicalized tokens are then looked up in the lexicon. Implemented lexicons are similar to their theoretical counterparts, comprising the word, its part of speech, and f-structure annotations such as PRED, CASE, and NUMBER. This information is integrated into the grammar via the annotated c-structure rules, as in theoretical LFG. Many implementations integrate a morphological analyzer which associates inflected forms of words with their lemma and morphological information. When using a morphological analyzer, the text is first tokenized and canonicalized for capitalization and then processed by the morphology. The output of the morphology (lemmata and morphological tags) are looked up in the lexicon. This simplifies the lexicon which only has to contain the lemmata and the morphological tags instead of containing all the inflected forms. These morphologies are often finite-state transducers (FSTs; Beesley & Karttunen 2003) which can be used for both parsing and generation.<sup>9</sup> For more details on using FSTs for preprocessing for LFG grammars see Kaplan et al. 2004 and Bögel et al. 2019, for integration of externally developed morphologies and lexicons within LFG grammars see Kaplan & Newman 1997.

A given inflected form can have multiple morphological analyses. Often all the analyses are provided as input to the LFG grammar and the c-structure rules

<sup>9</sup>Parsing goes from a string (e.g. a natural language sentence) to a c- and f-structure. Generation goes from an f-structure to a natural language string. Most theoretical LFG focuses on parsing, although some accounts, especially OT-LFG ones (see papers in Sells 2001), discuss generation.



4 Computational implementations and applications

Original text:	Dogs barked.		
Tokenization:	Dogs	barked	.
	dogs		
Morphology:	dog +Noun +Pl	bark +Verb +Past	. +Punct
	dog +Verb +3Sg		

Figure 1: Example preprocessing: Tokenization and morphological analysis

and f-structure constraints are used to eliminate analyses which are not feasible in the context of the sentence (e.g. the verbal analysis of *dogs* in figure 1). Preprocessing with a part-of-speech (PoS) tagger marks each word with its part of speech, as in (2). This information can be used to prune the morphological analyses and thus constrain the c-structure built over the sentence. Since PoS taggers are not perfect even for well-edited text, only certain tags are kept, or fall-back techniques are used when no analysis is found. See Kaplan & King 2003 and Dalrymple 2006 for more details on integrating PoS taggers into LFG and other symbolic grammars.

- (2) Dogs/Noun barked/Verb and/Conj the/Det cat/Noun left/Verb ./Punct

1.2.2 Projections

Theoretical LFG posits projections beyond the original lexicon, c-structure and f-structure. The exact number and combination of these projections is a subject of lively debate (Belyaev 2021 [this volume]). These include Lexical Mapping Theory (LMT) to map between underlying argument structure and grammatical functions in the lexicon (**chapters/Mapping**), phonological and prosodic projections (Bögel 2021 [this volume]), semantics and semantic structure (**chapters/Glue**), and information structure for discourse function information (**chapters/InformationStructure**). Most LFG implemented grammars do not include these additional projections because f-structures are sufficient for the applications they target. Even when other projections are included, they are often different from their theoretical counterparts both in their format and in how they are projected or derived. The primary additional component that is included is the semantic component. This component is based on the f-structure and is generally not a projection but instead is a separate post-processing step, although in some stages of its development the Norwegian ParGram grammar NorGram (Dyvik et al. 2016, 2019) included a semantic projection (Halvorsen 1983; Kaplan 1987; Halvorsen & Kaplan 1995) whose representations were in Minimal Recursion Semantics (Copestake et al. 2005).

*Martin Forst & Tracy Holloway King*

Semantic components include Glue semantics (Dalrymple et al. 1993; Meßmer & Zymła 2018) and ordered rewriting rules (Crouch & King 2006). The ordered rewriting rules have been extended into abstract knowledge representations (Bobrow et al. 2007). XLE-based implementations have been created for morphological structure (Butt et al. 1996) and prosodic structure and information structure (Butt & King 1998), although none of these are used in large-scale grammars. Instead, they focus on testing theoretical hypotheses and determining the complex interactions among different grammar components (Section 2). The lack of an implementation of LMT has resulted in issues for the parsing of morphologically rich languages like Turkish and Urdu, where interactions between passive and causative constructions cannot be easily captured in LFG implementations (Section 2; Çetinoğlu et al. 2009).

### 1.2.3 Ambiguity

Implemented grammars often include components to handle ambiguity (Section 1.5). There are three broad areas around managing ambiguity: computing all the analyses efficiently; representing the ambiguities compactly; resolving the ambiguity so that it does not need to be computed and represented. The first two are discussed in **chapters/Computational**. Within the grammar writer’s control are components including preprocessing by PoS taggers and named entity recognition systems, Optimality-Theory marks to prefer some constructions over others, and stochastic ranking of analyses.

### 1.2.4 Configuration

The determination of which components (e.g. which tokenizer, morphology, lexicons, and annotated c-structure rules) to use in an implemented grammar need to be specified in a configuration (see Crouch et al. 2008 on how this is done in XLE). These may have default values, e.g. a tokenizer which simply splits sentences at spaces and does not deal with capitalization or punctuation, but large-scale grammars require customized components for the specific language and often the type of text (e.g. newspaper text, tweets). In addition, to allow for rapid extension to specific applications which may have new vocabulary and unusual constructions, these configurations allow the grammar writer to specify lexicons and rules that add to or override those in a standard base grammar (King & Maxwell 2007). For example, to parse English academic biology papers, special lexicons of biological terms as well as special c-structure rules for section titles might be added to a grammar of standard written English.

## 4 Computational implementations and applications

### 1.3 Grammar Development Tools

To aid the grammar writer in managing a large-scale, broad-coverage LFG grammar, specialized variants of standard software development tools are needed. These grammar development tools are part of any LFG platform (Section 1.1). Throughout this chapter we rely on examples from XLE (Crouch et al. 2008) which is the most broadly adopted LFG grammar development framework and is used in the ParGram project (Section 3).

#### 1.3.1 Grammar writer interface

Grammar-development tools for the creation of LFG implementations facilitate the creation of c-structure rules and lexicon entries that are annotated with LFG functional annotations. Some platforms, e.g. Xerox’s Grammar Writer’s Workbench and XLFG, provide special interfaces for rules and lexicon entries. Others, e.g. XLE, use editors such as Emacs or the Eclipse-based eXLEpse (Rädle et al. 2011). The interfaces provide a way to apply the rules (i.e. the grammar) to a given input string and to output a c-structure and an f-structure graph in human-readable and machine-readable formats. They also generally provide tools to help debug issues such as why a well-formed input sentence does not receive an analysis or why the analysis is incorrect.

#### 1.3.2 Macros and templates

Since grammar engineers want to efficiently encode patterns across lexicon entries and grammar rules, some platforms support additional notations. XLE, for example, supports regular-expression macros that can expand to anything from a piece of f-annotation to an entire rule as well as f-annotation templates, e.g. to allow for like-category coordination over any c-structure category. Using a shared definition of templates across parallel LFG implementations for various languages and domains considerably facilitates the adherence to the agreed-upon f-structure conventions (King et al. 2005). For example, using a template `NUMBER` wherever number on nouns is assigned ensures that the same attribute (e.g. `NUMB`) is used and that it only needs to be changed in one place if later another name of the attribute is used (e.g. `NUM` instead of `NUMB`). See Section 2 for discussion of the role of macros and templates in theoretical LFG.

#### 1.3.3 Feature table and feature space

In a grammar formalism with untyped attribute-value matrices such as LFG, it is not strictly necessary to declare the valid values for the attributes used in

*Martin Forst & Tracy Holloway King*

f-structures and potential other levels of representation. However, from an engineering standpoint, it is highly desirable to make sure that only valid values are used; this way, unintended deviations due to typos can be caught easily (Crouch & King 2008). This need to enforce the adherence to a set of conventions is heightened in efforts to develop parallel LFG implementations for various languages such as ParGram (Section 3). XLE therefore supports feature declarations which state all the features, i.e. attributes, and their values that are allowed in the grammar. Multiple feature declarations can be combined to check the grammar code for adherence to them. In ParGram, each grammar combines the common feature declaration with a language-specific one which adds additional language-specific features and declares which subset of values are allowed, e.g. for English the *dual* value of the NUM attribute is removed.

#### 1.3.4 Treebanks as test suites:

Treebanks, and more specifically f-structure banks (**chapters/Treebanks**), can be used as a form of detailed, LFG-specific test suite for the grammar's coverage. Creating the treebank highlights missing constructions and vocabulary in the grammar. The grammar is then enhanced to account for these and the treebank is reparsed with the updated grammar and the new version of the treebank is inspected. This aids both in improving coverage and in ensuring that changes to the grammar do not break constructions that were previously covered. This approach has been used extensively in the development of the Norwegian (Dyvik et al. 2016), Polish (Patejuk & Przepiórkowski 2012), and Wolof (Dione 2014) grammars.

#### 1.3.5 Version control

Version control is used in software development to track changes to the software being developed. As with software development more generally, version control in grammar development allows the grammar writer to compare two versions of a rule, lexical entry, or any other part of the grammar, to revert to a previous version if needed, and to view conflicting changes. Version control systems also record who made a particular change, which makes it easier for multiple people to work on a grammar simultaneously by highlighting recent changes, especially conflicting ones. To our knowledge, eXLEPse (Rädle et al. 2011) is the only LFG-oriented editor that offers support for a variety of version-control systems. Since eXLEPse is based on Eclipse, all version-control plugins for Eclipse can be used. However, although XLE does not provide a version control system, most large

## 4 Computational implementations and applications

scale grammars use a standard software version control system such as SVN or Git. In addition, regression testing by providing sentences and analyses known to be parsable by the grammar help in determining whether new versions of a grammar function properly (Chatzichrisafis et al. 2007; de Paiva & King 2008).

### 1.3.6 Documentation

As with any software development project, it is important to document what each part of the implemented grammar does. This takes the form of comments in the lexicon and annotated phrase structure rules, including examples of sentences which that part of the grammar can parse. Dipper 2003 designed a self-documenting grammar system whereby the comments are extracted into proper, stand-alone documentation and example test suites of constructions covered by the grammar.

## 1.4 Modularity and Integration of Systems

LFG is an inherently modular linguistic theory, with different representations and components for the lexicon, phrase (constituent) structure, functional structure, semantics, etc. This modularity is highlighted in implemented systems which introduce two other types of modularity: modularity for the grammar components, which correlates with the linguistic modularity, and modularity within those components, which enables better grammar engineering practices. LFG implementations are software systems and hence modularity of the different components is important for developing, scaling, maintaining and debugging the system. This section describes how the modularity of the grammar components helps with grammar implementation.

A core tenet of LFG is that different parts of the grammar require different types of representations. This is echoed in the implementations where the different modules can be created by different people and use different types of technology. As with theoretical LFG, the c-structure is a tree and the f-structure an attribute-value matrix, and the two are related via annotated phrase-structure rules. These phrase-structure rules form one module of the grammar. Similarly, lexicons comprise word forms, parts-of-speech, and f-annotations. These form another module. These lexicons can be custom-created for the LFG grammar or converted from other lexical resources (Kaplan & Newman 1997; Sheil & Ørsnes 2006; Przepiórkowski et al. 2014; Patejuk & Przepiórkowski 2014). The morphological component is often implemented as a finite-state transducer (Kaplan et al.

*Martin Forst & Tracy Holloway King*

2004; Bögel et al. 2019) but can be of any form.<sup>10</sup> For example, the ParGram Chinese grammar uses a combined tokenizer and part-of-speech tagger that was externally developed for non-LFG purposes (Fang & King 2007). The importance of modularity is highlighted by the treatment of semantics: there have been many implementational approaches to semantic representations based on the LFG f-structure analyses. These include projecting the semantics as an attribute-value matrix (Halvorsen 1983; Halvorsen & Kaplan 1995; Asudeh 2006; Dyvik et al. 2016, 2019), implementing Glue Semantics (Dalrymple et al. 1993; Meßmer & Zymła 2018), and using ordered rewrite rules (Crouch & King 2006). Without a modular system, this exploration of the best way to capture the semantics would be difficult.

There are three additional reasons to maintain modularity in an implemented grammar. The first is that large scale grammars often have multiple grammar writers. By having different files for the lexicon, templates, and annotated phrase structure rules, the efforts can be divided in such a way that changes can be easily merged. To further aid this, the lexicons and phrase-structure rules often comprise multiple files, e.g. the lexicon might be divided into verbs, closed-class items, and all other entries, and the phrase-structure rules might be divided into clausal and nominal. The second reason is that debugging, i.e. the process of finding and fixing errors in the grammar, is simpler in a more modular system. By having different components and different files within those components, the structure of the grammar is easier to see and the individual rules easier to locate. This debugging is further aided by the use of test suites (Chatzichrisafis et al. 2007; de Paiva & King 2008), including ones based on examples in comments in the grammar rules (Dipper 2003). Even with modularity, the inclusion of OT marks (Section 1.5) can make debugging more complex since an analysis may not surface due to competition with another analysis. A third reason is that as described in Section 1.2 and Section 1.3, in addition to a lexicon and annotated phrase structure rules, LFG implementations can have tokenizers, morphologies, templates, feature tables, etc. These are combined via configuration files that encode the different modules of the system and the way they interact.

## 1.5 Runtime Performance

When implemented grammars are used to test linguistic hypotheses and analyses (Section 2), how quickly the grammar provides an analysis for a sentence, i.e. its latency, is generally not important. However, almost all other uses for implemented grammars (Section 4) have latency considerations. LFG implementations

---

<sup>10</sup>The non-FST morphologies are referred to as library transducers in XLE.

#### 4 Computational implementations and applications

have provided a number of techniques to improve latency, sometimes at the cost of accuracy and coverage, e.g. certain analyses may be lost due to early elimination of possible structures (Kaplan et al. 2004). There are two main issues with runtime performance of LFG grammars: ambiguity and latency. These considerations hold for both parsing and generation; we focus on parsing here.<sup>11</sup>

Ambiguity concerns the multiple analyses (i.e. c- and f-structures) that are assigned to a given sentence. The ambiguity problem is accentuated when there is no semantic or pragmatic processing to guide the choice among the different analyses. The ambiguities fall into three broad categories. First, sentences can have multiple analyses, all of which are correct and equally plausible out of context, e.g. in *I saw her duck* either I saw a bird or I saw a person ducking down. Second, sentences can have correct analyses but even out of context some of them are highly improbable, e.g. in *I saw the child with the telescope* there are two plausible readings where *saw* is the past tense of the verb *see* and one implausible one where *saw* is the present tense of the verb meaning to cut with a saw, which is only plausible in a bizarre magic show. Third, ambiguities can arise when the grammar allows ungrammatical analyses, either intentionally as a fall-back mechanism or unintentionally due to an error in the implementation. Copperman & Segond 1996 provide one of the first detailed expositions of ambiguity in LFG grammars, comparing the ambiguity discussed in the theoretical linguistics literature with that in implemented grammars. King et al. 2004 discuss ambiguity in LFG grammar writing in detail, focusing on the XLE-based LFG implementations.

Language contains ambiguities at many levels from determining word boundaries in tokenization, to morphological analysis, to syntactic attachment ambiguities, to semantic quantifier scope and beyond. This can result in thousands of analyses even for short sentences and long processing times to compute each analysis. There are two main ways to handle this ambiguity efficiently. One is to handle the ambiguity by “packing” (Maxwell & Kaplan 1989, 1993; Shemtov 1997) and operating at each level efficiently over the packed representations. Packing allows operations to apply just once to shared parts of the representation instead of enumerating all of the possibilities and processing each of them separately. For example, XLE is designed to maintain packed structures from the tokenization and morphology to the syntactic c- and f-structures and then into an ordered rule writing system that can be used to create semantic representations (Crouch & King 2006). The other way to handle ambiguity is to choose the most likely analysis at each level. For example, if there are multiple morphological analyses

---

<sup>11</sup>See **chapters/Computational** on the inherent formal and computational properties of LFG.



*Martin Forst & Tracy Holloway King*

for a word (e.g. English *leaves*), the system can choose the most likely one given the information it has at that time (e.g. the words adjacent to *leaves* and their potential morphological analyses). This has the downside that the correct analysis may be lost due to removing information early (Dalrymple 2006).

Optimality Theory (OT) (chapters/OT) can be used to allow the grammar writer to prefer certain analyses and even to control which grammar rules are active. Frank et al. 1998, 2001 propose an extension of the classical LFG projection architecture to incorporate a constraint ranking mechanism inspired by OT. A new projection, the o-projection, specifies violable constraints, which are used to determine a “winner” among competing, alternative analyses. Many ambiguities can be filtered from the set of possible analyses for a given sentence by using this constraint ranking mechanism in the XLE system. For example, OT marks can be used to prefer verbal analyses over adjectival ones in copular clauses with passives like *They were eaten*. XLE further provides a way to cut down the search space in parsing, allowing for potentially fewer parses to search through. This is done via a special STOPPOINT feature, which is part of the Optimality Theory preference mechanism incorporated into XLE (King et al. 2000). The OT marks can be grouped with certain groups only applying if no parse is found with the original set of OT marks. That is, XLE will process the input in multiple passes, using larger and larger versions of the grammar in subsequent reparsing phases. These groupings are referred to as STOPPOINTS. STOPPOINTS are useful for eliminating ungrammatical analyses when grammatical analyses are present and for speeding up the parser by only using expensive and rare constructions when no other analysis is available. If a solution can be found with the smaller, restricted grammar, XLE will terminate with this solution. Otherwise, a reparsing phase is triggered. This approach can be used to prefer multi-word expressions, for instance so that XLE will only consider analyses that involve the individual components of the multi-word expression if there is no valid analysis involving the multi-word expression. In addition to the OT marks, c-structure pruning (Cahill et al. 2008) and part-of-speech tagging and named entity recognition (Kaplan & King 2003; Dalrymple 2006; Krasnowska-Kieraś & Patejuk 2015) can be used to eliminate unlikely c-structures before unification.

Even with the use of OT marks, a sentence may have many valid parses. However, downstream applications often expect a single analysis, i.e. a single f-structure, as input. To use LFG grammars as input to such applications, statistical methods can be used to choose the most probable analysis (Riezler et al. 2002). These stochastic models are trained on treebanks or dependency banks of known correct analyses. As a variant of this, Dalrymple 2006 and Krasnowska-Kieraś &



## 4 *Computational implementations and applications*

Patejuk 2015 investigated using a stochastic part-of-speech tagger to trim potential analyses before constructing the c- and f-structure.

## 2 Implications for theoretical issues

LFG and HPSG (Bender & Emerson 2019) are in the privileged position of having not only a community of theoretical linguists but also of grammar engineers, with significant crossover between the theoretical and grammar-engineering communities. There are four areas in which grammar engineering interacts with theoretical linguistics (King 2011, 2016). These include: using grammar engineering to confirm linguistic hypotheses; linguistic issues highlighted by grammar engineering; implementation capabilities guiding theoretical analyses; and insights into architecture issues. The positive feedback loop between theoretical and implementational efforts is a domain in which LFG and HPSG have a distinct advantage compared to many other linguistic theories, given the strong communities and resources available.

### 2.1 Confirming Linguistic Hypotheses

Grammar engineering can be used to confirm linguistic hypotheses (Bierwisch 1963; Müller 1999; Butt, Dipper, et al. 1999; Bender 2008; Bender et al. 2011; King 2011; Fokkens 2014; King 2016). Encoding the hypothesis in an implemented grammar not only highlights details of the analysis that might be missed in a pencil-and-paper version but can also bring to light interesting interactions with other linguistic phenomena, especially when the hypothesized analysis is encoded in a broad-coverage grammar. Two examples of this type include the analysis of determiner agreement systems and the prosody-syntax interaction.

King & Dalrymple 2004 provide an LFG analysis of determiner agreement and noun conjunction, looking particularly at indeterminacy of agreement features. In order to test the proposed system, they implemented a toy grammar with lexical entries of each type and enough syntactic structure to encompass determiner, adjective, and verb agreement with conjoined and non-conjoined nouns. As a result, the authors were able to confirm that their analysis was formally sound and accounted for the known data. This toy grammar was relatively easy to implement in XLE because all of the necessary components, e.g. distributive features, were already available.

Implementing proposals for the prosody-syntax interaction in LFG are more challenging because not all of the mechanisms that have been proposed in the

*Martin Forst & Tracy Holloway King*

literature are available in systems like XLE. Butt & King 1998 used an existing, non-LFG analysis of Bengali clitics and implemented it in order to test whether p(rosodic)-structure could be used to capture the generalizations proposed in the theoretical analysis, focusing on where mismatches between prosodic and syntactic structure occur. A much different interface approach was pursued in Bögel et al. 2009, which built upon the finite-state transducers used for tokenization and morphological analysis within the grammars (Section 1.2). Finally, a large-scale implementation of certain phonology-syntax interactions was completed for Welsh (Mittendorf & Sadler 2006).

## 2.2 Implementational Devices

Writing large-scale grammars highlights the interaction of different parts of the grammar and the need to be able to formally state certain types of generalizations. These needs have led to the creation of formal devices, some of which have become part of theoretical LFG analyses while others remain implementational devices. Implementation capabilities that guided theoretical analysis include the use of complex categories for auxiliary analysis in English and German, the analysis of Welsh phonology-syntax interactions through the interaction of morphological analysis via finite-state transducers and the LFG c-structure, and the introduction of templates and macros.

Complex categories (Crouch et al. 2008) are a formal c-structure device. They allow for generalizations over c-structure categories by having the category be composed of a fixed component and a variable, where the variable can pass its value to other complex categories on the right-hand side of the rule. In this way, they allow the grammar writer to capture generalizations through notation. This notation is then automatically compiled into standard c-structure rules. Complex categories are used to constrain the order and form of auxiliaries and main verbs in English (e.g. *They will have been promoted.*) by having each auxiliary state its meaning and its form (e.g. *have* is an AUX[perf,base] with perfective meaning and base form while *been* is an AUX[pass,perf] with passive meaning and perfective form) and the VP rules themselves are complex categories that reflect their head and based on that put requirements on their complement.

Welsh consonant mutations are a phenomenon whereby the initial consonant of certain words changes based on its phonological and syntactic environment (**chapters/Celtic**). To capture the joint requirements on the morphophonology and the syntax which trigger mutations, Mittendorf & Sadler 2006 used the finite-state morphology capabilities integrated in XLE to control where Welsh consonant mutations occur by encoding the boundary conditions in the morphological

## 4 Computational implementations and applications

tag sequences. The modular nature of LFG combined with the implementational device of finite-state morphology provided a clean solution to the different types of triggers for the mutations.

A long standing debate in the linguistic literature, especially for constraint-based formalisms like HPSG and LFG, is whether a comprehensive and efficient grammatical theory should include a type hierarchy and what role it should play. Historically HPSG has had types as foundational to the theory while LFG has not. However, in grammar engineering, it is important to be able to efficiently capture generalizations as well as exceptions to those generalizations. The introduction of templates into the formal devices available to LFG allows for generalizations and inheritance via notation, without introducing a full type hierarchy into the formalism (Dalrymple, Kaplan & King 2004; Crouch & King 2008) and as a result, the concept of templates has become part of theoretical LFG analyses. Similar to complex categories, templates and macros allow the grammar writer to capture generalizations through notation, which is then automatically compiled into standard LFG c- and f-structure rules.

Two more minor formal devices which are gaining traction in theoretical analyses are instantiation and local variables (a third is the restriction operator discussed in the next section). Since the beginning, predicates (PRED) in LFG have not been unifiable with one another due to their unique lexical index (Kaplan & Bresnan 1982). Certain non-PRED features also need to be non-unifiable (Dalrymple 2001). This can be captured by instantiation, represented by having the value of the feature be followed by an underscore. For example, instantiating the form values of English particles blocks their occurring multiple times in a sentence (e.g. *\*they threw out the garbage out*) (see Figure 2 for an English example and Forst et al. 2010). Finally, local variables anchor a functional uncertainty equation to a particular f-structure and then refer to that f-structure in other annotations (Dalrymple 2001; Crouch et al. 2008). This is needed when making a set of statements about a particular element of a set or a particular type of governing element. For example Szűcs 2019 uses local variables to state constraints on topic left dislocation constructions in Hungarian.

### 2.3 Architectural Issues

Implementing a wide variety of phenomena, as is necessary for broad-coverage grammars, brings to light architectural issues with the theory. Çetinoğlu et al. 2009 and Bögel et al. 2019 describe issues with the interaction of the passive and causative in Turkish and Urdu. These issues are the result of how lexical rules in

*Martin Forst & Tracy Holloway King*

LFG interact with complex predicate formation, where the passive is traditionally analyzed as involving a lexical rule while the causative is often analyzed as a complex predicate. The Urdu and Turkish grammars use the restriction operator (Kaplan & Wedekind 1993) in the annotated c-structure rules to model complex predication, including causatives. The restriction operator allows for features of f-structures to be restricted out, i.e. to cause the grammar to function as if these features did not exist. This allows complex predicate-argument structures to be built dynamically (Butt et al. 2003, 2010). In contrast, the passive is handled by lexical rules which apply to the predication frames in the lexicon. This predicts that passivization applies before causativization and that it is not possible to passivize a causative by demoting or suppressing the subject of the causative. However, this is the reverse of the Urdu and Turkish facts. To solve this problem in the ParGram grammars of Urdu and Turkish, both the causative and the passive are handled via restriction in the annotated phrase structure rules. In the theoretical literature, this issue had not been highlighted because for Turkish and Urdu style morphosyntax, the causative was handled in argument-structure, but the interaction between causativization and passives at the morphology-syntax interface highlighted that traditional lexical rules do not allow for the right order of application when causativization is morphological but passivization is part of the syntax.

To conclude this section, the interaction of grammar engineering and theoretical linguistics helps to confirm linguistic hypotheses, to highlight complex linguistic issues, to posit new formal capabilities, and shed light on architecture issues. The positive feedback loop between theoretical and implementational efforts is a domain in which LFG and HPSG have a distinct advantage.

### **3 Grammar Resources: ParGram**

The systems described above are used to create small- and large-scale LFG grammars. These can be used as input to applications (Section 4) or to explore theoretical hypotheses (Section 2). The Parallel Grammar (ParGram) project is a consortium of LFG researchers implementing grammars for a typologically varied set of languages in a parallel fashion (Butt, King, Niño, et al. 1999; Butt et al. 2002) using the XLE LFG parser, generator, and grammar development platform. The parallels are most notable in the f-structure space, where common features and analyses are used wherever possible, but differ when required by the syntax of the languages. This parallelism is enabled by LFG theory, by grammar engineering components such as feature declarations, and by semi-annual meetings

#### 4 Computational implementations and applications

between the grammar writers.<sup>12</sup>

ParGram began with three languages: English (Riezler et al. 2002), French (Frank 1996), and German (Dipper 2003; Rohrer & Forst 2006). They developed aligned f-structure analyses for a tractor manual which existed as an aligned corpus in all three languages. Even with three closely related languages, it was clear that full f-structure alignment was not possible (Butt, Dipper, et al. 1999) due to fundamental syntactic differences in the languages. Later, the Fuji Xerox Corporate Research Group and the University of Bergen joined the initiative with a Japanese (Masuichi et al. 2003) and a Norwegian grammar (Dyvik et al. 2016, 2019) respectively. Other longer-term academic efforts participating in ParGram concern the development of Urdu (Butt & King 2002, 2007) and Polish (Patejuk & Przepiórkowski 2012) LFG implementations. Finally, further ParGram efforts have given rise to computational LFGs for Arabic (Attia 2006, 2012), Chinese (Fang & King 2007), Danish (Ørsnes 2006), Georgian (Meurer 2009), Hungarian (Laczkó & Rákosi 2008–2019), Indonesian (Arka et al. 2009; Arka 2012), Korean (Kim et al. 2003), Malagasy (Dalrymple et al. 2006), Tamil (Sarveswaran & Butt 2019) Tigrinya (Kifle 2011), Turkish (Çetinoğlu & Oflazer 2018), Welsh (Mittendorf & Sadler 2006), and Wolof (Dione 2014).

The project resulted in the creation of LFG grammars in these multiple languages and hence a greater understanding of the parallelism (or lack thereof) for the LFG analyses of particular constructions. Major issues in LFG analysis and architecture highlighted by the ParGram project included: Copular constructions and in particular whether there is a copular *be* predicate and whether the predicated argument has a subject (xCOMP-like) or not (PREDLINK) (Dalrymple, Dyvik, et al. 2004; Attia 2008); how to handle argument-changing relations such as the passive, causative, benefactives, complex predicates, and interactions thereof, including morphological and syntactic interactions (Bögel et al. 2019; see Section 2); whether auxiliaries have predicates or just supply tense and aspect features to the f-structure (Butt et al. 1996; Dyvik 1999); the interaction of tokenization and morphology with the c- and f-structures, especially around features like Welsh mutations (Mittendorf & Sadler 2006) and Urdu complex predicates (Bögel et al. 2019). In addition, the ParGram project resulted in improvements to the grammar development platform (Section 1.2, Section 1.3 and Section 1.5) and in best practices for distributed parallel grammar development.

In addition to the traditional LFG-style ParGram grammars which use annotated phrase structure rules to create the c- and f-structure representations,

---

<sup>12</sup>A similar approach was subsequently adopted by the HPSG DELPH-IN consortium (Bender et al. 2002).

*Martin Forst & Tracy Holloway King*

the ParGram project also includes several automatically induced grammars that create ParGram compatible f-structures, i.e. f-structures using the same feature space as the grammars described above, but which are learned from tree and f-structure banks (Cahill et al. 2002). These grammars are robust in that they produce f-structures for nearly any sentence, at the cost of producing structures which sometimes violate core LFG principles such as completeness and coherence. See Section 4 for applications which require such robustness.

An influential initiative that resembles ParGram is the Universal Dependencies (UD) initiative (McDonald et al. 2013; see also **chapters/Dependency**). Like ParGram, it aims at parallel representations across languages, and UD follows LFG concerning many of the distinctions made at the level of syntactic dependencies and grammatical functions respectively (de Marneffe et al. 2014). This being said, surface-oriented dependency structures as used in UD cannot be as parallel as the more abstract f-structures of ParGram. Korsak 2018 and Przepiórkowski & Patejuk 2020 discuss the similarities between LFG and UD and investigate mapping between LFG f-structures and UD. Another noteworthy difference between ParGram and UD is that ParGram has been developing reversible XLE grammars whereas UD focuses solely on parsing.

## 4 Applications

Some applications integrating natural language processing only require parsing. For these applications, parsing should be robust to typos and grammatical errors, unusual constructions, unknown words, etc. In addition, minor issues in parsing may be unimportant for these applications because systematic errors can be compensated for within the system. Semantic search is an application that requires only parsing, needs to be robust, and can tolerate certain parsing errors.

Other applications, e.g. sentence condensation, transfer-based machine translation (MT) and conversational agents, require both parsing and generation. Applications using generation generally require highly grammatical output since users are sensitive to malformed natural language such as incorrect subject-verb agreement. Since corpus-induced grammars do not lend themselves to refinement in order to control generation, hand-crafted grammar implementations such as LFG grammars are still the means of choice for the generation of high-quality text.

Finally, there are applications that require grammaticality judgments. This is the case of grammar checkers, both general-purpose ones and grammar checkers for computer-assisted language learning (CALL). Parsers trained on general-

## 4 Computational implementations and applications

purpose treebanks cannot be used for this purpose, so these applications are another natural fit for hand-crafted grammar implementations. In our opinion, LFG suits this purpose particularly well because its terminology is relatively close to that used in language instruction.

### 4.1 Applications requiring deep features and robustness

For applications that require mainly natural language understanding, parsing needs to be robust to unexpected words and constructions. To provide the robustness necessary for these applications, domain-specific grammars can be created based on a general large-scale grammar (Kim et al. 2003; King & Maxwell 2007). However, this is often not enough to cover all use cases. LFG grammars can use morphological guessers to cover unknown vocabulary (Dost & King 2009; Bögel et al. 2019), can parse fragments of the structure, e.g. provide f-structures for all the noun phrases even if they cannot be formed into a sentence (Riezler et al. 2003), and can include fall-back rules (mal-rules Schneider & McCoy 1998; Reuer 2003; Khader 2003; Fortmann & Forst 2004; Bender et al. 2004) explicitly accounting for certain types of ungrammaticality, e.g. incorrect subject-verb agreement.

Semantic search is one application which benefits from the deep LFG representations. As a search application, the goal is to find documents which are relevant to the query and, ideally, to highlight the passage in the document most relevant to the query. Semantic search moves beyond keyword matching to match the relationships between entities in the query. It can include queries that are full interrogatives as well as ones that are phrases. The ParGram XLE English grammar was used in the Powerset Inc. semantic search engine for searching Wikipedia articles. By using LFG representations for the query and the documents it can differentiate between *who acquired PeopleSoft* and *who did PeopleSoft acquire*, where PeopleSoft is the object in the first question and the subject in the second. By using a fragment grammar as a backup, longer sentences could be partially parsed, e.g. the first conjunct of a coordinated sentence could be parsed even if the second failed. This combined with the redundancy across the articles made using an LFG grammar feasible for moving beyond keyword search. The f-structures were mapped to abstract knowledge representations which went beyond grammatical functions to semantic rules, e.g. mapping *Oracle acquired PeopleSoft* and *PeopleSoft was acquired by Oracle* and even *Oracle's acquisition of PeopleSoft* to the same abstract representation.

A more complex application than semantic search is question answering. Unlike search, question answering uses a document collection to find the answer to the query, which is generally in the form of a natural-language question, and



*Martin Forst & Tracy Holloway King*

present it to the user. The PARC Bridge system (Bobrow et al. 2007) used the XLE ParGram English grammar as its base and mapped the query and documents to an abstract knowledge representation using ordered rewrite rules, deep lexical resources such as WordNet (Fellbaum 1998) and VerbNet (Kipper et al. 2000; Levin 1993), and knowledge resources such as Cyc (Lenat 1995). The queries and documents were then matched against one another with a graph-based algorithm. An interesting extension of this was to perform entailment and contradiction detection (ECD) (Bobrow et al. 2007) with a graph-based module that determined whether one sentence entailed or contradicted (or neither) the other. ECD depended on understanding the roles between the entities as determined by the LFG grammar as well as detailed lexical knowledge.

Burton 2006 describes a tutorial system which uses the XLE English grammar for its language-understanding component. The tutorial system is provided by Acuitus and teaches network administration. The coursework includes a set of troubleshooting exercises where students find and fix problems. During these exercises the computer helps the students when they ask for help or based on their actions. The system asks the student a mix of multiple-choice, short-answer, and natural-language questions. The idea behind using natural-language interactions is to encourage students to think beyond what multiple-choice questions provide and to allow more complex questions and answers. The system converts the f-structures from the student input to semantic interpretations via the transfer rule system (Crouch 2006). Both the syntactic parsing and the semantics are adapted to the domain to provide more accurate and robust results.

Historically, hand-crafted LFG implementations have had a hard time competing with machine-learned constituency or dependency parsers in terms of robustness, i.e. providing a parse for all input, and speed for purely understanding-oriented applications, even though they are often superior in terms of systematicity and detail of analysis and despite the fact that machine-learned parsers often produce illogical parses for input where LFG grammars would fail to produce a parse. Because of this speed and perceived robustness, machine-learning-based dependency parsers have become increasingly popular, as is evident from the shared tasks of the Conference on Computational Natural Language Learning (CoNLL) series. Interesting though, the CoNLL tasks now often integrate UD representations (McDonald et al. 2013), which can be seen as less fine-grained f-structures (see Section 3 for more details on UD). The combination of hand-crafted grammars, fall-back techniques, and statistical parser selection as described in this chapter allow LFG and other rule-based grammars to be used in applications requiring robustness (see also Ivanova et al. 2106).



## 4 Computational implementations and applications

### 4.2 Applications requiring grammaticality

Certain applications not only aim to map text to representations more amenable to the computation of meaning, but they also take abstract meaning representations, including f-structures, as input and map them to text. Among such applications are sentence condensation and transfer-based machine translation, both applications for which LFG implementations have been used because f-structures are abstract enough to facilitate transformations like the removal of certain adjuncts or the transfer from a source to a target language. Furthermore, since corpus-induced grammars do not lend themselves to refinement in order to control generation, hand-crafted grammar implementations are still the means of choice for the generation of high-quality text.

Sentence condensation is a form of summarization (Knight & Marcu 2000; Jing 2000). It takes a long sentence and produces a shorter sentence which preserves the core meaning of the original sentence. This requires the ability to identify the core part of the original sentence and to generate a grammatical shorter sentence. Riezler et al. 2003 and Crouch et al. 2004 used the ParGram XLE grammar to create a sentence condensation system for English. The LFG f-structure was used to identify the core meaning, e.g. by removing adjuncts other than negation. A new f-structure was created which contained only this core meaning. This new f-structure was then run through the grammar in the generation direction to generate the shorter, condensed sentence. This sentence was guaranteed to be grammatical since it met the well-formedness conditions of the grammar. Since multiple strings (e.g. sentences) can map to the same f-structure, more than one condensed sentence can often be generated from a single f-structure. This can be partially controlled by Optimality-Theory marks in the grammar in XLE (Frank et al. 1998). The choice between the remaining sentences can be done with a language model (Riezler et al. 2003). A related application is note taking where longer texts are condensed into legible notes (Kaplan et al. 2005).

Machine translation (MT) involves automatically translating a text from one language (the source) to another (the target). The resulting translation has to preserve the meaning and to be grammatical. LFG f-structures have been used for MT (Oepen et al. 2004; Riezler & Maxwell 2006; Avramidis & Kuhn 2009; Graham et al. 2009; Graham 2012; Graham & van Genabith 2012; Homola & Coler 2012). The idea is that the f-structure encodes the meaning of the sentence more abstractly than the surface form of the text and so can be used as the level for translation. That is, f-structure enables translation by transfer across structures and not just an interlingua across words (Kaplan et al. 1989). In theory, simply substituting the PRED values in the f-structure could produce an f-structure in the

*Martin Forst & Tracy Holloway King*

target language and the LFG grammar can then be used to generate the translation. In practice, f-structures still encode enough language-specific syntactic information that additional transfer rules need to be applied before the generation step. For example, one language may use indefinite singular determiners (e.g. English *a*) while the other may not, in which case the determiner would have to be deleted (in the source language) or inserted (in the target language). The LOGON MT project (Oepen et al. 2004) provides an interesting approach with parsing via the LFG Norwegian NorGram grammar, transfer to semantic MRS (Copestake et al. 2005) and generation via an HPSG English grammar. Although LFG-based MT systems can be brittle since there has to be a successful parse, transfer, and generation, when a translation is produced it is generally of high quality both in terms of preserving the meaning and of being grammatically well-formed.

Consider the English and German sentences in (3) and (4), for which the corresponding f-structures are displayed in Figure 2.

- (3) Across the city, monuments to prosperity have sprung up.
- (4) In der ganzen Stadt sind Denkmäler des Wohlstands entstanden.  
in the whole city be monuments of.the prosperity up.spring  
Across the city, monuments to prosperity have sprung up.

Apart from the fact that the German analysis of adjunct NPs in the genitive is not parallel to other ParGram implementations and that the German finite-state morphology decomposes the word *Wohlstand*, which gives rise to a MOD dependency under the SUBJ ADJ-GEN, the f-structures are surprisingly parallel. (At first sight, this is obscured by the fact that in the German f-structure, the sub-f-structures under TOPIC and in the ADJUNCT set are the same.) Even though the English sentence is headed by a particle verb while the German one is not, there is a single PRED value for the head verb on either side; even though the subject of the English sentence precedes the verb while the one of the German sentence follows the verb, both appear in the respective f-structure under SUBJ; even though the auxiliary in the English sentence is *have* while the German verb *entstehen* requires the auxiliary *sein* ('to be') for perfect tenses, the auxiliaries contribute the same value for TNS-ASP PERF. As a result, the transfer component can concentrate on word-to-word translation equivalencies while letting the language-specific grammars take care of well-formedness conditions independent of the language pair under consideration. An example of a non-trivial translation equivalency is the one between *across the city* and *in der ganzen Stadt* (literally 'in the entire city'), as the English phrase might also correspond to *durch die Stadt* (literally 'through the city') in other contexts (especially in combination with motion verbs).

#### 4 Computational implementations and applications

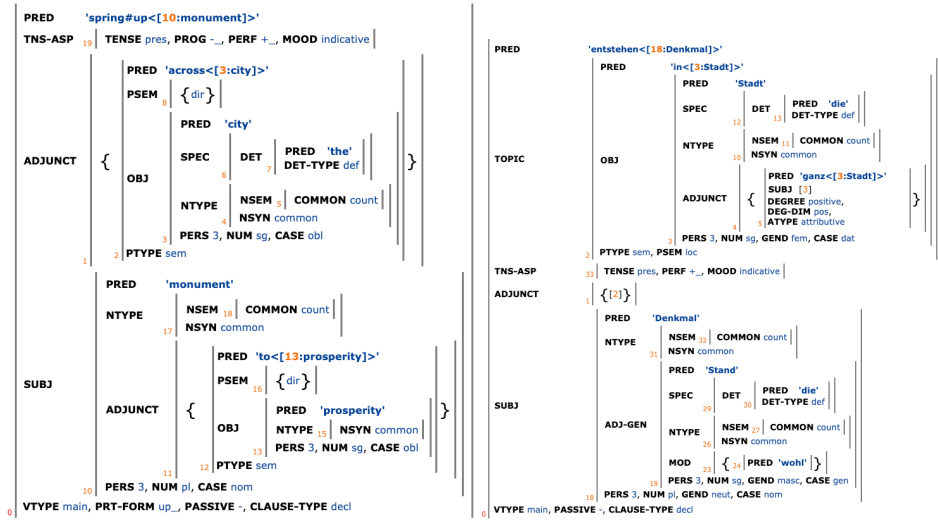


Figure 2: F-structures for English and German translation equivalents

Certain other applications do not require semantic representations or grammatical text output but do require the system to have a notion of grammaticality as their purpose is to highlight ungrammatical (or otherwise undesired) passages in text. Such systems can be directed to a general public of people producing texts or explicitly target second-language learners, sometimes even second-language learners with a specific first-language background. The latter application, in the context of Intelligent Computer-Assisted Language Learning (ICALL), has used LFG implementations, typically augmented with mal-rules (Rypa & Feuerman 1995; Reuer 2003; Khader 2003; Fortmann & Forst 2004). Mal-rules are rules or rule extensions that cover ungrammatical constructions typically produced by second-language learners, e.g. NPs where determiners or adjectives do not agree with the head noun, NPs with countable head nouns in the singular that are not preceded by a determiner, or sentences with an ungrammatical order of constituents or a violation of subject-verb agreement. As typical mistakes made by second-language learners depend significantly on their native language as well as on other languages they know, malrules can be optimized with respect to their coverage more easily when the linguistic background of the audience is known. A machine-learning-based approach to ICALL exploiting features provided by the English ParGram LFG implementation is described by Berend et al. 2013.

A final application we discuss are natural language understanding (NLU) components used in car computers or in personal assistants on mobile devices. Those

*Martin Forst & Tracy Holloway King*

NLU components often combine grammar-based analysis and deep-learning-based neural networks or statistical models learned from annotated data. Moreover, machine-learning-based NLU models depend on large amounts of training data from the relevant domain. Since such data is hard to collect and costly to annotate, much of it is generated by means of grammars. For the most part, the grammars used to this end are simple, largely context-free grammars. However, as the semantic representations used for NLU become increasingly sophisticated, the use of more powerful grammar formalisms such as LFG can be used for the generation of high-quality grammatical training data.

## 5 Conclusion

This chapter provided an overview of computational implementations of LFG. LFG was designed from the outset to be computationally tractable and has a strong history of broad-coverage implementations for multiple languages, primarily through the ParGram project which is built on the XLE grammar development platform. As with theoretical LFG, implemented grammars primarily focus on c-structure and f-structure, but extensive work has been done on using the resulting f-structures as input to semantics and abstract knowledge representation, and some work has focused on the integration of morphological and phonological information as well as argument structure. The ParGram project is based on the theoretical LFG hypothesis that languages are more similar at f-structure, which encodes grammatical functions, than at c-structure. This f-structure similarity can then be exploited in applications such as machine translation. Other applications which take advantage of the more abstract f-structures and the ability of LFG grammars to parse and generate as well as to detect (un)grammaticality include computer-assisted language learning, question answering, and sentence condensation. From a theoretical linguistic perspective, implemented grammars allow the linguist to test analyses and to see interactions between different parts of the grammar.

## 6 Acknowledgements

We would like to thank Emily Bender, Gerlof Bouma, Ron Kaplan, Agnieszka Patejuk, Annie Zaenen and two anonymous reviewers for detailed comments on this chapter. All remaining errors are our own.

## References

- Arka, I Wayan. 2012. Developing a deep grammar of Indonesian within the Par-Gram framework: Theoretical and implementational challenges. In *Proceedings of the 26th Pacific Asia conference on language, information and computation*, 19–38.
- Arka, I Wayan, Avery D. Andrews, Mary Dalrymple, Meladel Mistica & Jane Simpson. 2009. A linguistic and computational morphosyntactic analysis for the applicative *-i* in Indonesian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 85–105. Stanford, CA: CSLI Publications.
- Asudeh, Ash. 2006. Direct compositionality and the architecture of LFG. In Miriam Butt, Mary Dalrymple & Tracy Holloway King (eds.), *Intelligent linguistic architectures: Variations on themes by Ronald M. Kaplan*, 363–387. Stanford, CA: CSLI Publications.
- Attia, Mohammed. 2006. Accommodating multiword expressions in an Arabic LFG grammar. In *Advances in natural language processing (FinTAL 2006)*, vol. 4139 (Lecture Notes in Computer Science), 87–98. DOI: 10.1007/11816508\_11.
- Attia, Mohammed. 2008. A unified analysis of copula constructions in LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 89–108. Stanford, CA: CSLI Publications.
- Attia, Mohammed. 2012. *Ambiguity in Arabic computational morphology and syntax: A study within the Lexical Functional Grammar framework*. Saarbrücken: LAP LAMBERT Academic Publishing.
- Avramidis, Eleftherios & Jonas Kuhn. 2009. Exploiting XLE's finite state interface in LFG-based statistical machine translation. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 127–145. Stanford, CA: CSLI Publications.
- Beesley, Kenneth R. & Lauri Karttunen. 2003. *Finite state morphology*. Stanford, CA: CSLI Publications.
- Belyaev, Oleg. 2021. Introduction to LFG. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 3–20. Berlin: Language Science Press. DOI: ??.
- Bender, Emily M. 2008. Grammar engineering for linguistic hypothesis testing. In Nicholas Gaylord, Alexis Palmer & Elias Ponvert (eds.), *Proceedings of the Texas Linguistics Society X conference: Computational linguistics for less-studied languages*, 16–36. Stanford, CA: CSLI Publications.
- Bender, Emily M. & Guy Emerson. 2019. Computational linguistics and grammar engineering. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre

*Martin Forst & Tracy Holloway King*

- Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Berlin: Language Science Press.
- Bender, Emily M., Dan Flickinger & Stephan Oepen. 2002. The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk & Richard Sutcliffe (eds.), *COLING-GEE '02: Proceedings of the 2002 workshop on Grammar Engineering and Evaluation*, 8–14. DOI: 10.3115/1118783.1118785.
- Bender, Emily M., Dan Flickinger & Stephan Oepen. 2011. Grammar engineering and linguistic hypothesis testing: Computational support for complexity in syntactic analysis. In Emily M. Bender & Jennifer E. Arnold (eds.), *Language from a cognitive perspective: Grammar, usage and processing*, 5–29. Stanford, CA: CSLI Publications.
- Bender, Emily M., Dan Flickinger, Stephan Oepen, Annemarie Walsh & Tim Baldwin. 2004. Arboretum: Using a precision grammar for grammar checking in CALL. In *Proceedings of the InSTIL/ICALL symposium: NLP and speech technologies in advanced language learning systems*. Venice.
- Berend, Gabor, Veronika Vincze, Sina Zarriß & Richárd Farkas. 2013. LFG-based features for noun number and article grammatical errors. In *Proceedings of the 17th Conference on Computational Natural Language Learning: Shared task*, 62–67.
- Bierwisch, Manfred. 1963. *Grammatik des deutschen Verbs*. Vol. II (Studia Grammatica). Berlin: Akademie Verlag.
- Bobrow, Daniel G., Bob Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, Valeria de Paiva, Charlotte Price & Annie Zaenen. 2007. PARC’s bridge and question answering system. In Tracy Holloway King & Emily M. Bender (eds.), *Proceedings of the GEAF07 workshop*. Stanford, CA: CSLI Publications. <http://csli-publications.stanford.edu/GEAF/2007/geaf07-toc.html>.
- Bögel, Tina. 2021. Prosody and its interfaces. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 95–137. Berlin: Language Science Press. DOI: ??.
- 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, CA: CSLI Publications.
- Bögel, Tina, Miriam Butt & Tracy Holloway King. 2019. Urdu morphology and beyond: Why grammars should not live without finite-state methods. In Cleo

#### 4 Computational implementations and applications

- Condoravdi & Tracy Holloway King (eds.), *Tokens of meaning: Papers in honor of Lauri Karttunen*, 417–438. Stanford, CA: CSLI Publications.
- Boullier, Pierre & Benoît Sagot. 2005. Efficient and robust LFG parsing: SxLFG. In *Proceedings of the 9th International Workshop on Parsing Technologies (IWPT 2005)* (Parsing '05), 1–10. Stroudsburg, PA: Association for Computational Linguistics. DOI: 10.3115/1654494.1654495.
- Burton, Richard R. 2006. Using XLE in an intelligent tutoring system. In Miriam Butt, Mary Dalrymple & Tracy Holloway King (eds.), *Intelligent linguistic architectures: Variations on themes by Ronald M. Kaplan*, 75–90. Stanford, CA: CSLI Publications.
- Butt, Miriam, Stefanie Dipper, Anette Frank & Tracy Holloway King. 1999. Writing large-scale parallel grammars for English, French and German. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '99 conference*. Stanford, CA: CSLI Publications.
- Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi & Christian Rohrer. 2002. The Parallel Grammar Project. In John Carroll, Nelleke Oostdijk & Richard Sutcliffe (eds.), *COLING-GEE '02: Proceedings of the 2002 workshop on Grammar Engineering and Evaluation*, 1–7. Taipei: Association for Computational Linguistics. DOI: 10.3115/1118783.1118786.
- Butt, Miriam & Tracy Holloway King. 1998. Interfacing phonology with LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '98 conference*. Stanford, CA: CSLI Publications.
- Butt, Miriam & Tracy Holloway King. 2002. Urdu and the Parallel Grammar project. In N. Calzolari, K.-S. Choi, A. Kawtrakul, A. Lenci & T. Takenobu (eds.), *Proceedings of the 3rd workshop on Asian language resources and international standardization, 19th International Conference on Computational Linguistics (COLING '02)*, 39–45. DOI: 10.3115/1118759.1118762.
- Butt, Miriam & Tracy Holloway King. 2007. Urdu in a parallel grammar development environment. *Language Resources and Evaluation* 41(2). Special Issue on Asian Language Processing: State of the Art Resources and Processing, 191–207. DOI: 10.1007/s10579-007-9042-8.
- Butt, Miriam, Tracy Holloway King & John T. Maxwell III. 2003. Complex predication via restriction. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '03 conference*, 92–104. Stanford, CA: CSLI Publications.
- Butt, Miriam, Tracy Holloway King, María-Eugenia Niño & Frédérique Segond. 1999. *A grammar writer's cookbook*. Stanford, CA: CSLI Publications.
- Butt, Miriam, Tracy Holloway King & Gillian Ramchand. 2010. Complex predication: Who made the child pinch the elephant? In Linda Ann Uyechi & Lian-Hee



*Martin Forst & Tracy Holloway King*

- Wee (eds.), *Reality exploration and discovery: Pattern interaction in language and life*, 231–256. Stanford, CA: CSLI Publications.
- Butt, Miriam, María-Eugenia Niño & Frederique Segond. 1996. Multilingual processing of auxiliaries in LFG. In D. Gibbon (ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS conference*, 111–122. Berlin: Mouton de Gruyter.
- Cahill, Aoife & Martin Forst. 2009. Human evaluation of a German surface realisation ranker. In *Proceedings of the 12th conference of the European chapter of the ACL (EACL 2009)*, 112–120. Association for Computational Linguistics. DOI: 10.3115/1609067.1609079.
- Cahill, Aoife, John T. Maxwell III, Paul Meurer, Christian Rohrer & Victoria Rosén. 2008. Speeding up LFG parsing using c-structure pruning. In *COLING 2008: Proceedings of the workshop on Grammar Engineering Across Frameworks*, 33–40. Manchester. DOI: 10.3115/1611546.1611551.
- Cahill, Aoife, Mairéad McCarthy, Josef van Genabith & Andy Way. 2002. Parsing with PCFGs and automatic f-structure annotation. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '02 conference*, 76–95. Stanford, CA: CSLI Publications.
- Čavar, Damir, Lwin Moe, Hai Hu & Kenneth Steimel. 2016. Preliminary results from the Free Linguistic Environment project. In Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King & Stefan Müller (eds.), *Proceedings of the joint 2016 conference on Head-Driven Phrase Structure Grammar and Lexical Functional Grammar*, 161–181. Stanford, CA: CSLI Publications.
- Çetinoğlu, Özlem, Miriam Butt & Kemal Oflazer. 2009. Mono/bi-clausality of Turkish causatives. In Sila Ay, Özgür Aydın, İclâl Ergenç, Seda Gökmen, Selçuk İşsever & Dilek Peçenek (eds.), *Essays on Turkish linguistics: Proceedings of the 14th international conference on Turkish linguistics*. Wiesbaden: Harrassowitz Verlag.
- Çetinoğlu, Özlem & Kemal Oflazer. 2018. Deep parsing of Turkish with Lexical-Functional Grammar. In Kemal Oflazer & Murat Saraçlar (eds.), *Turkish natural language processing (Theory and Applications of Natural Language Processing)*, 175–206. Springer. DOI: 10.1007/978-3-319-90165-7\_9.
- Chatzichrisafis, Nikos, Richard Crouch, Tracy Holloway King, Rowan Nairn, Manny Rayner & Marianne Santaholma. 2007. Regression testing for grammar-based systems. In Tracy Holloway King & Emily M. Bender (eds.), *Proceedings of the GEAF07 workshop*, 128–143. Stanford, CA: CSLI Publications. <http://csli-publications.stanford.edu/GEAF/2007/geaf07-toc.html>.



## 4 Computational implementations and applications

- Clément, Lionel. 2019. Une étude de la coordination des propositions avec ellipse en français : Formalisation et application avec XLFG. *Langue française* 203. 35–52. DOI: 10.3917/lf.203.0035.
- Clément, Lionel & Alexandra Kinyon. 2001. XLFG: An LFG parsing scheme for French. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '01 conference*. Stanford, CA: CSLI Publications.
- Copestake, Ann, Dan Flickinger, Ivan A. Sag & Carl Pollard. 2005. Minimal Recursion Semantics: An introduction. *Research on Language and Computation* 3. 281–332. DOI: 10.1007/s11168-006-6327-9.
- Copperman, Max & Frederique Segond. 1996. Computational grammars and ambiguity: The bare bones of the situation. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '96 conference*. Stanford, CA: CSLI Publications.
- Crouch, Richard. 2006. Packed rewriting for mapping text to semantics and KR. In Miriam Butt, Mary Dalrymple & Tracy Holloway King (eds.), *Intelligent linguistic architectures: Variations on themes by Ronald M. Kaplan*. Stanford, CA: CSLI Publications.
- Crouch, Richard, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III & Paula Newman. 2008. *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).
- Crouch, Richard & Tracy Holloway King. 2006. Semantics via f-structure rewriting. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '06 conference*, 145–165. Stanford, CA: CSLI Publications.
- Crouch, Richard & Tracy Holloway King. 2008. Type-checking in formally non-typed systems. In *Proceedings of the ACL workshop on software engineering, testing, and quality assurance for natural language processing*, 3–4. DOI: 10.3115/1622110.1622112.
- Crouch, Richard, Tracy Holloway King, John T. Maxwell III, Stefan Riezler & Annie Zaenen. 2004. Exploiting f-structure input for sentence condensation. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*, 167–187. Stanford, CA: CSLI Publications.
- Dalrymple, Mary. 2001. *Lexical Functional Grammar*. Vol. 34 (Syntax and Semantics). New York: Academic Press. DOI: 10.1163/9781849500104.
- Dalrymple, Mary. 2006. How much can part-of-speech tagging help parsing? *Natural Language Engineering* 12. 373–389. DOI: 10.1017/s1351324905004079.
- Dalrymple, Mary, Helge Dyvik & Tracy Holloway King. 2004. Copular complements: Closed or open? In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*, 188–198. Stanford, CA: CSLI Publications.

*Martin Forst & Tracy Holloway King*

- Dalrymple, Mary, Ronald M. Kaplan & Tracy Holloway King. 2004. Linguistic generalizations over descriptions. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*, 199–208. Stanford, CA: CSLI Publications.
- Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III & Annie Zaenen (eds.). 1995. *Formal issues in Lexical-Functional Grammar*. Stanford, CA: CSLI Publications.
- Dalrymple, Mary, John Lamping & Vijay Saraswat. 1993. LFG semantics via constraints. In *Proceedings of the 6th conference of the European chapter of the ACL (EACL 1993)*, 97–105. DOI: [10.3115/976744.976757](https://doi.org/10.3115/976744.976757).
- Dalrymple, Mary, Maria Liakata & Lisa Mackie. 2006. Tokenization and morphological analysis for Malagasy. *International Journal of Computational Linguistics & Chinese Language Processing* 11(4). 315–332.
- de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre & Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik.
- de Paiva, Valeria & Tracy Holloway King. 2008. Designing testsuites for grammar-based systems in applications. In *COLING 2008: Proceedings of the workshop on Grammar Engineering Across Frameworks*, 49–56. DOI: [10.3115/1611546.1611553](https://doi.org/10.3115/1611546.1611553).
- Dione, Cheikh M. Bamba. 2014. LFG parse disambiguation for Wolof. *Journal of Language Modelling* 2(1). DOI: [10.15398/jlm.v2i1.81](https://doi.org/10.15398/jlm.v2i1.81).
- Dipper, Stefanie. 2003. Implementing and documenting large-scale grammars – German LFG. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)* 9(1).
- Dost, Ascander & Tracy Holloway King. 2009. Using large-scale parser output to guide grammar development. In Tracy Holloway King & Marianne Santaholma (eds.), *Proceedings of the 2009 workshop on Grammar Engineering Across Frameworks (GEAF 2009)*, 63–70. Association for Computational Linguistics. DOI: [10.3115/1690359.1690367](https://doi.org/10.3115/1690359.1690367).
- Dyvik, Helge. 1999. The universality of f-structure: Discovery or stipulation? The case of modals. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '99 conference*, 1–11. Stanford, CA: CSLI Publications.
- Dyvik, Helge, Gyri Smørdal Losnegaard & Victoria Rosén. 2019. Multiword expressions in an LFG grammar for Norwegian. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions*. Berlin: Language Science Press. DOI: [10.5281/zenodo.2579037](https://doi.org/10.5281/zenodo.2579037).

#### 4 Computational implementations and applications

- Dyvik, Helge, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse & Martha Thunes. 2016. NorGramBank: A ‘deep’ treebank for Norwegian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC’16)*, 3555–3562. Portorož, Slovenia. <http://www.lrec-conf.org/proceedings/lrec2016/summaries/943.html>.
- Earley, Jay. 1970. An efficient context-free parsing algorithm. *Communications of the ACM* 13(2). 94–102. DOI: 10.1145/362007.362035.
- Fang, Ji & Tracy Holloway King. 2007. An LFG Chinese grammar for machine use. In Tracy Holloway King & Emily M. Bender (eds.), *Proceedings of the GEAF07 workshop*, 144–160. Stanford, CA: CSLI Publications. <http://csli-publications.stanford.edu/GEAF/2007/geaf07-toc.html>.
- Fellbaum, Christiane (ed.). 1998. *Wordnet: An electronic lexical database*. Cambridge, MA: The MIT Press. DOI: 10.7551/mitpress/7287.001.0001.
- Fokkens, Antske Sibelle. 2014. *Enhancing empirical research for linguistically motivated precision grammars*. Saarbrücken: Universität des Saarlandes. (Doctoral dissertation).
- Forst, Martin. 2003. Treebank conversion – Establishing a testsuite for a broad-coverage LFG from the TIGER treebank. In *Proceedings of 4th international workshop on linguistically interpreted corpora (LINC-03) at EACL 2003*, 205–216. <https://www.aclweb.org/anthology/W03-2404>.
- Forst, Martin. 2007. Disambiguation for a linguistically precise German parser. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)* 13(3).
- Forst, Martin, Tracy Holloway King & Tibor Laczkó. 2010. Particle verbs in computational LFGs: Issues from English, German, and Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’10 conference*, 228–248. Stanford, CA: CSLI Publications.
- Fortmann, Christian & Martin Forst. 2004. An LFG grammar checker for CALL. In *Proceedings of the InSTIL/ICALL symposium: NLP and speech technologies in advanced language learning systems*.
- Frank, Anette. 1996. A note on complex predicate formation: Evidence from auxiliary selection, reflexivization, passivization and past participle agreement in French and Italian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’96 conference*. Stanford, CA: CSLI Publications.
- Frank, Anette, Tracy Holloway King, Jonas Kuhn & John T. Maxwell III. 1998. Optimality Theory style constraint ranking in large-scale LFG grammars. In

*Martin Forst & Tracy Holloway King*

- Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '98 conference*, 1–16. Stanford, CA: CSLI Publications.
- 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, CA: CSLI Publications.
- Graham, Yvette. 2012. Deep syntax in statistical machine translation. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 240–253. Stanford, CA: CSLI Publications.
- Graham, Yvette, Anton Bryl & Josef van Genabith. 2009. F-structure transfer-based statistical machine translation. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 317–337. Stanford, CA: CSLI Publications.
- Graham, Yvette & Josef van Genabith. 2012. Exploring the parameter space in statistical machine translation via f-structure transfer. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 254–270. Stanford, CA: CSLI Publications.
- Halvorsen, Per-Kristian. 1983. Semantics for Lexical-Functional Grammar. *Linguistic Inquiry* 14. 567–615.
- Halvorsen, Per-Kristian & Ronald M. Kaplan. 1995. Projections and semantic description in Lexical-Functional Grammar. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III & Annie Zaenen (eds.), *Formal issues in Lexical-Functional Grammar*, 279–292. Stanford, CA: CSLI Publications.
- Homola, Petr & Matt Coler. 2012. Machine translation using dependency representation. Presented at the LFG '12 Conference.
- Ivanova, Angelina, Stephan Oepen, Rebecca Drīdan, Dan Flickinger, Lilja Øvrelid & Emanuele Lapponi. 2106. On different approaches to syntactic analysis into bi-lexical dependencies: An empirical comparison of direct, PCFG-based, and HPSG-based parsers. *Journal of Language Modelling* 4(1). 113–144. DOI: 10.15398/jlm.v4i1.101.
- Jing, Hongyan. 2000. Sentence reduction for automatic text summarization. In *Proceedings of the 6th applied natural language processing conference (ANLP'00)*. DOI: 10.3115/974147.974190.
- Kaplan, Ronald M. 1987. Three seductions of computational psycholinguistics. In Peter Whitelock, Mary McGee Wood, Harold L. Somers, Rod Johnson & Paul Bennett (eds.), *Linguistic theory and computer applications*, 149–188. London: Academic Press. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 339–367).

## 4 Computational implementations and applications

- Kaplan, Ronald M. 2009. Deep natural language processing for web-scale search. Presented at the LFG '09 Conference.
- Kaplan, Ronald M. & Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 173–281. Cambridge, MA: The MIT Press. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 29–130).
- Kaplan, Ronald M., Richard Crouch, Tracy Holloway King, Michael Tepper & Danny Bobrow. 2005. A note-taking appliance for intelligence analysts. In *Proceedings of the international conference on intelligence analysis*.
- Kaplan, Ronald M. & Tracy Holloway King. 2003. Low-level markup and large-scale LFG grammar processing. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '03 conference*, 238–249. Stanford, CA: CSLI Publications.
- Kaplan, Ronald M. & John T. Maxwell III. 1996. *LFG Grammar Writer's Workbench*. Xerox Palo Alto Research Center. Palo Alto, CA. [https://www.researchgate.net/profile/John\\_Maxwell5/publication/2760068\\_Grammar\\_Writer's\\_Workbench/links/0c96052405e97928e9000000.pdf](https://www.researchgate.net/profile/John_Maxwell5/publication/2760068_Grammar_Writer's_Workbench/links/0c96052405e97928e9000000.pdf).
- Kaplan, Ronald M., John T. Maxwell III, Tracy Holloway King & Richard Crouch. 2004. Integrating finite-state technology with deep LFG grammars. In *Proceedings of the workshop on combining shallow and deep processing for NLP at the European summer school on logic, language, and information (ESSLI)*.
- Kaplan, Ronald M., Klaus Netter, Jürgen Wedekind & Annie Zaenen. 1989. Translation by structural correspondences. In *Proceedings of the 4th conference of the European chapter of the ACL (EACL 1989)*, 272–281. DOI: 10.3115/976815.976852. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 311–330).
- Kaplan, Ronald M. & Paula S. Newman. 1997. Lexical resource reconciliation in the Xerox Linguistic Environment. In *Proceedings of the ACL workshop on computational environments for grammar development and engineering*.
- Kaplan, Ronald M. & Jürgen Wedekind. 1993. Restriction and correspondence-based translation. In *Proceedings of the 6th conference of the European chapter of the ACL (EACL 1993)*, 193–202. DOI: 10.3115/976744.976768.
- Khader, I. R. M. A. K. 2003. *Evaluation of an English LFG-based grammar as error checker*. Manchester: University of Manchester. (M.Sc. thesis).
- Kifle, Nazareth Amlesom. 2011. *Tigrinya applicatives in Lexical-Functional Grammar*. Bergen: University of Bergen. (Doctoral dissertation).
- Kim, Roger, Mary Dalrymple, Ron Kaplan & Tracy Holloway King. 2003. Multilingual grammar development via grammar porting. In *Proceedings of the 17th Pacific Asia conference on language, information and computation*, 98–105.

*Martin Forst & Tracy Holloway King*

- King, Tracy Holloway. 2011. (Xx<sup>\*</sup>-)linguistics: Because we love language. *Linguistic Issues in Language Technology: Interaction of Linguistics and Computational Linguistics* 6.
- King, Tracy Holloway. 2016. Theoretical linguistics and grammar engineering as mutually constraining disciplines. In Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King & Stefan Müller (eds.), *Proceedings of the joint 2016 conference on Head-Driven Phrase Structure Grammar and Lexical Functional Grammar*, 339–359. Stanford, CA: CSLI Publications.
- King, Tracy Holloway & Mary Dalrymple. 2004. Determiner agreement and noun conjunction. *Journal of Linguistics* 4. 69–104. DOI: [10.1017/s0022226703002330](https://doi.org/10.1017/s0022226703002330).
- 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 (ESSLLI-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](https://doi.org/10.1023/b:rolc.0000016784.26446.98).
- King, Tracy Holloway, Martin Forst, Jonas Kuhn & Miriam Butt. 2005. The feature space in parallel grammar writing. *Research on Language and Computation* 3(2). 139–163. DOI: [10.1007/s11168-005-1295-z](https://doi.org/10.1007/s11168-005-1295-z).
- King, Tracy Holloway & John T. Maxwell III. 2007. Overlay mechanisms for multi-level deep processing applications. In Tracy Holloway King & Emily M. Bender (eds.), *Proceedings of the GEAF07 workshop*, 182–2002. Stanford, CA: CSLI Publications. <http://csli-publications.stanford.edu/GEAF/2007/geaf07-toc.html>.
- Kipper, Karin, Hoa Trang Dang & Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the 17th national conference on artificial intelligence (AAAI-2000)*.
- Knight, Kevin & Daniel Marcu. 2000. Statistics-based summarization – step one: Sentence compression. In *Proceedings of the 17th national conference on artificial intelligence (AAAI-2000)*.
- Korsak, Katarzyna Magdalena. 2018. *LFG-based universal dependencies for Norwegian*. Oslo: University of Oslo. (MA thesis).
- Krasnowska-Kieraś, Katarzyna & Agnieszka Patejuk. 2015. Integrating Polish LFG with external morphology. In Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk & Adam Przepiórkowski (eds.), *Proceedings of the 14th international workshop on treebanks and linguistic theories (TLT14)*, 134–147.
- Laczko, Tibor & György Rákosi. 2008–2019. *HunGram: An XLE implementation*. Tech. rep. Debrecen: University of Debrecen.



#### 4 Computational implementations and applications

- Lenat, Doug. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38(11). DOI: 10.1145/219717.219745.
- Levin, Beth. 1993. *English verb classes and alternations*. Chicago: University of Chicago Press.
- Masuichi, Hiroshi, Tomoko Ohkuma, Hiroki Yoshimura & Yasunari Harada. 2003. Japanese parser on the basis of the Lexical-Functional Grammar formalism and its evaluation. In *Proceedings of the 17th Pacific Asia conference on language, information and computation*.
- Maxwell, John T., III. 2006. Efficient generation from packed input. In Miriam Butt, Mary Dalrymple & Tracy Holloway King (eds.), *Intelligent linguistic architectures: Variations on themes by Ronald M. Kaplan*, 19–34. Stanford, CA: CSLI Publications.
- 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. 1993. The interface between phrasal and functional constraints. *Computational Linguistics* 19. 571–590.
- McDonald, Ryan, Joakim Nivre, Yvonne Quirnbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló & Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (ACL’13)*, 92–97.
- Meßmer, Moritz & Mark-Matthias Zymla. 2018. The glue semantics workbench: A modular toolkit for exploring linear logic and glue semantics. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’18 conference*, 268–282. Stanford, CA: CSLI Publications.
- Meurer, Paul. 2009. A computational grammar for Georgian. In Peter Bosch, David Gabelaia & Jérôme Lang (eds.), *Logic, language, and computation: 7th international Tbilisi symposium on logic, language, and computation, TbiLLC*, vol. 5422. Springer. DOI: 10.1007/978-3-642-00665-4\_1.
- Minos, Panagiotis. 2014. *Development of a natural language parser for the LFG formalism*. Athens: National & Kapodistrian University of Athens. (MA thesis).
- Mittendorf, Ingo & Louisa Sadler. 2006. A treatment of Welsh initial mutation. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’06 conference*. Stanford, CA: CSLI Publications.

Martin Forst & Tracy Holloway King

- Müller, Stefan. 1999. *Deutsche Syntax deklarativ: Head-Driven Phrase Structure Grammar für das Deutsche*. Tübingen: Max Niemeyer Verlag. DOI: 10.1515/9783110915990.
- Oepen, Stephan, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård & Victoria Rosén. 2004. *Som å kapp-ete med trollet? Towards MRS-based Norwegian-English Machine Translation*. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*.
- Ørnes, Bjarne. 2006. Creating raising verbs: An LFG-analysis of the complex passive in Danish. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '06 conference*. Stanford, CA: CSLI Publications.
- Patejuk, Agnieszka & Adam Przepiórkowski. 2012. Towards an LFG parser for Polish: An exercise in parasitic grammar development. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, 3849–3852. European Language Resources Association (ELRA).
- Patejuk, Agnieszka & Adam Przepiórkowski. 2014. Synergistic development of grammatical resources: A valence dictionary, an LFG grammar, and an LFG structure bank for Polish. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, 113–126. Department of Linguistics (SfS), University of Tübingen.
- Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski & Marek Świdziński. 2014. Walenty: Towards a comprehensive valence dictionary of Polish. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, 2785–2792.
- Przepiórkowski, Adam & Agnieszka Patejuk. 2020. From Lexical Functional Grammar to Enhanced Universal Dependencies: The UD-LFG treebank of Polish. *Language Resources and Evaluation* 54. 185–221. DOI: 10.1007/s10579-018-9433-z.
- Rädle, Roman, Michael Zöllner & Sebastian Sulger. 2011. eXLEPse: An Eclipse-based, easy-to-use editor for computational LFG grammars. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '11 conference*, 422–439. Stanford, CA: CSLI Publications.
- Reuer, Veit. 2003. Error recognition and feedback with Lexical Functional Grammar. *CALICO Journal* 20. 497–512.



#### 4 Computational implementations and applications

- Riezler, Stefan, Tracy Holloway King, Richard Crouch & Annie Zaenen. 2003. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for Lexical-Functional Grammar. In *Proceedings of the 2003 Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics*, 197–204. DOI: 10.3115/1073445.1073471.
- 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 (ACL'02)*. Philadelphia.
- Riezler, Stefan & John T. Maxwell III. 2006. Grammatical machine translation. In *Proceedings of the human language technology conference of the NAACL, main conference*, 248–255. New York City: Association for Computational Linguistics. DOI: 10.3115/1220835.1220867.
- Rohrer, Christian & Martin Forst. 2006. Improving coverage and parsing quality of a large-scale LFG for German. In Miriam Butt, Mary Dalrymple & Tracy Holloway King (eds.), *Intelligent linguistic architectures: Variations on themes by Ronald M. Kaplan*. Stanford, CA: CSLI Publications.
- Rosén, Victoria, Koenraad De Smedt, Paul Meurer & Helge Dyvik. 2012. An open infrastructure for advanced treebanking. In Jan Hajič, Koenraad De Smedt, Marko Tadić & António Branco (eds.), *Proceedings of the META-research workshop on advanced treebanking at LREC'12*, 22–29. Istanbul: European Language Resources Association (ELRA).
- Rosén, Victoria, Paul Meurer & Koenraad De Smedt. 2009. LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In Frank Van Eynde, Anette Frank, Gertjan van Noord & Koenraad De Smedt (eds.), *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT7)*, 127–133. Utrecht: LOT.
- Rypa, Marikka & Ken Feuerman. 1995. CALLE: An exploratory environment for foreign language learning. In V. Holland, J. Kaplan & M. Sams (eds.), *Intelligent language tutors: Theory shaping technology*, 55–76. Lawrence Erlbaum Associates.
- Sagot, Benoît & Pierre Boullier. 2006. Deep non-probabilistic parsing of large corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*. Genoa: European Language Resources Association (ELRA). [http://www.lrec-conf.org/proceedings/lrec2006/pdf/806\\_pdf.pdf](http://www.lrec-conf.org/proceedings/lrec2006/pdf/806_pdf.pdf).

*Martin Forst & Tracy Holloway King*

- Sagot, Benoît, Lionel Clément, Éric De La Clergerie & Pierre Boullier. 2006. The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*.
- Sarveswaran, Kengatharaiyer & Miriam Butt. 2019. Computational challenges with Tamil complex predicates. In Miriam Butt, Tracy Holloway King & Ida Toivonen (eds.), *Proceedings of the LFG '19 conference*, 272–292. Stanford, CA: CSLI Publications.
- Schneider, David & Kathleen F. McCoy. 1998. Recognizing syntactic errors in the writing of second language learners. In *ACL '98/COLING '98: Proceedings of the 36th annual meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 1198–1204. Montréal: Association for Computational Linguistics. DOI: 10.3115/980691.980765.
- Sells, Peter (ed.). 2001. *Formal and empirical issues in optimality theoretic syntax*. Stanford, CA: CSLI Publications.
- Sheil, Beau & Bjarne Ørsnes. 2006. Using a large external dictionary in an LFG grammar: The STO experiments. In Miriam Butt, Mary Dalrymple & Tracy Holloway King (eds.), *Intelligent linguistic architectures: Variations on themes by Ronald M. Kaplan*, 167–198. Stanford, CA: CSLI Publications.
- Shemtov, Hadar. 1997. *Ambiguity management in natural language generation*. Stanford, CA: Stanford University. (Doctoral dissertation).
- Szűcs, Péter. 2019. Left dislocation in Hungarian. In Miriam Butt, Tracy Holloway King & Ida Toivonen (eds.), *Proceedings of the LFG '19 conference*, 293–313. Stanford, CA: CSLI Publications.
- Tomita, Masaru (ed.). 1991. *Current issues in parsing technology*. Dordrecht: Kluwer Academic Publishers. DOI: 10.1007/978-1-4615-3986-5.
- Zweigenbaum, Pierre. 1991. Un analyseur pour grammaires lexicales-fonctionnelles. *TA Informations* 32. 19–34.

## Chapter 5

# Treebank-driven Parsing, Translation and Grammar Induction using LFG

Aoife Cahill

Educational Testing Service

Andy Way

ADAPT Centre, School of Computing, Dublin City University

This chapter provides a summary of a range of work on probabilistic models of Lexical Functional Grammar (LFG). LFG grammars as originally conceived in Kaplan & Bresnan (1982) were defined by grammatical rules and constraints, so could not describe ill-formed strings, and they failed if confronted with well-formed strings outside their coverage. In contrast, the hybrid LFG-DOP model of Bod & Kaplan 1998 and Bod & Kaplan 2003 could generalize well-formed analyses via the *Discard* operation to allow ill-formed and previously uncovered well-formed strings to be handled.

Way (1999) and Way (2001) extended LFG-DOP to handle translation, and demonstrated two advantages of his LFG-DOT models: (i) being probabilistic, LFG-DOT was able to handle a range of translation phenomena that were problematic for the description of LFG-MT (Kaplan et al. 1989); and (ii) having f-structure constraints enabled LFG-DOT to overcome problems for DOT (Poutsma 2000), a model of translation based on DOP (Bod 1992; Sima'an 1997; Bod 1998).

Like most probabilistic models, LFG-DOP (and LFG-DOT) require large amounts of annotated data. In a range of seminal work on grammar induction – now a research field in its own right, but at the time quite a novelty – it was demonstrated how strings could be automatically annotated with both LFG c- and f-structure information (Sadler et al. 2000; Cahill et al. 2002a). These were then used for multilingual probabilistic parsing (Cahill et al. 2005; Cahill, Burke, O'Donovan, Riezler, et al. 2008) and lexicon induction experiments (O'Donovan 2006), which we describe here.



*Aoife Cahill & Andy Way*

## 1 Introduction

In this chapter we summarize work on extensions to the core LFG formalism that facilitate large-scale probabilistic LFG parsing and translation models. Traditional LFG grammars (Kaplan & Bresnan 1982) are defined in terms of well-formed grammatical rules and constraints. This has two main limitations: (i) ill-formed input cannot be handled easily;<sup>1</sup> and (ii) when a grammar produces multiple analyses for an input, there is no inherent way of ranking the competing solutions.

We describe LFG-DOP (Bod & Kaplan 1998), a hybrid model of Data-Oriented Parsing (DOP: Bod 1992; Sima'an 1997; Bod 1998) and LFG that allows for probabilistic tree parsing, and which is beyond context-free in its generative power. We describe how this work led to the LFG-DOT framework (Way 1999, 2001) for machine translation (MT) with LFG.

Large-scale probabilistic parsing typically requires substantial amounts of annotated training data. We describe techniques developed to automatically generate large-scale LFG-annotated treebanks that provide the training data needed for probabilistic LFG parsing. We describe how this work was not only applied to English, but also several other languages including German (Cahill et al. 2005; Rehbein & van Genabith 2009), French (Schluter 2011), Spanish (O'Donovan et al. 2005; Chrupała & van Genabith 2006), Chinese (Burke, Cahill, et al. 2004; Guo 2009), Japanese (Oya & van Genabith 2007) and Arabic (Tounsi et al. 2009a). A related field of work was the automated extraction of large-scale lexical resources from these LFG-annotated treebanks (O'Donovan 2006). Although large-scale LFG-DOT experimentation has not been conducted to date,<sup>2</sup> these grammars and semantic forms (i.e. subcategorisation frames) are exactly what LFG-DOT requires to build its models. Accordingly, we sketch what would need to be done to conduct such experiments.

Finally, we compare this semi-automatic approach to lexicon and grammar induction to that based on the hand-crafted XLE grammars.

---

<sup>1</sup>This applies equally to legitimate strings which are not covered by the grammar.

<sup>2</sup>Bod (2000) acknowledges that Cormons (1999) “accomplished [the] first simple experiment with LFG-DOP”. Bod & Kaplan (2003) includes a large-scale evaluation of LFG-DOP against a DOP baseline. Hearne (2005) extends these experiments for DOP, demonstrating higher accuracy for the exact match metric using improved sampling techniques.

## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

### 2 LFG-DOP

This section describes how LFG was combined with Data-Oriented Parsing (DOP) models to create a more robust, probabilistic model of language processing, LFG-DOP. In later sections, we will show how both DOP and LFG-DOP were used to build powerful, robust models of MT.

#### 2.1 Data-Oriented Parsing

DOP models (Bod 1992; Sima'an 1997; Bod 1998) assume that past experiences of language are significant in both perception and production. DOP prefers performance models over competence grammars, in that abstract grammar rules are eschewed in favour of models based on large collections of previously occurring fragments of language. Previously uncovered sentences are processed with reference to existing fragments from the treebank, which are combined using probabilistic techniques to determine the most likely analysis for the new fragment.

The general DOP architecture stipulates four parameters on which particular models are instantiated:

1. A formal definition of *well-formed representations* for sentence analyses;
2. A set of *decomposition* operations for splitting sentence analyses into a set of fragments;
3. A set of *composition* operations for recombination of such fragments in order to derive analyses of new strings;
4. A definition of a *probability model* indicating the likelihood of a sentence analysis based on the probabilities of its constituent parts.

DOP models typically assign a surface phrase-structure (PS) tree to strings (hence ‘Tree-DOP’, or ‘DOP1’ in Bod (1992)). However, context-free models are insufficiently powerful to deal with all aspects of human language. LFG, on the other hand, is known to be beyond context-free, and can capture and provide representations of linguistic phenomena other than those occurring at surface structure.<sup>3</sup>

---

<sup>3</sup>Note that the question of what grammar type in the Chomsky Hierarchy (Chomsky 1956) was capable of processing human language was a significant one when LFG was first proposed, but appears to be less of a concern nowadays. This was relevant for Chomsky’s claims of Universal Grammar (Chomsky 1981), of course, but different languages have been demonstrated to require different grammar types; for example, Dutch cross-serial dependencies can only be

*Aoife Cahill & Andy Way*

## 2.2 Combining DOP with LFG: LFG-DOP

Accordingly, Bod & Kaplan (1998) augmented DOP with the syntactic representations of LFG to create a new, more powerful hybrid model of language processing – LFG-DOP – which adds a level of robustness not available to models based solely on LFG.

LFG-DOP is defined using the same four parameters as in Tree-DOP. We describe each of these in the next sections.

### 2.2.1 Representations in LFG-DOP

The LFG-DOP **representations** are those traditionally used in LFG, where each string is annotated with a c-structure, an f-structure, and a mapping  $\phi$  between them. Well-formedness conditions operate solely on f-structure, as usual.

### 2.2.2 Decomposition in LFG-DOP

Since we are now dealing with  $\langle c, f \rangle$  pairs of structure, the *Root* and *Frontier decomposition* operations of DOP need to be adapted to stipulate exactly which c-structure nodes are linked to which f-structure fragments, thereby maintaining the fundamentals of c- and f-structure correspondence. As LFG c-structures are little more than annotated PS trees, we can proceed very much on the same lines as in Tree-DOP. *Root* erases all nodes outside of the selected node, and in addition deletes all  $\phi$ -links (informally, parts of the f-structure linked to a c-structure node) leaving the erased nodes, as well as all f-structure units that are not  $\phi$ -accessible from the remaining nodes. Bod & Kaplan (1998) define  $\phi$ -accessibility as follows:

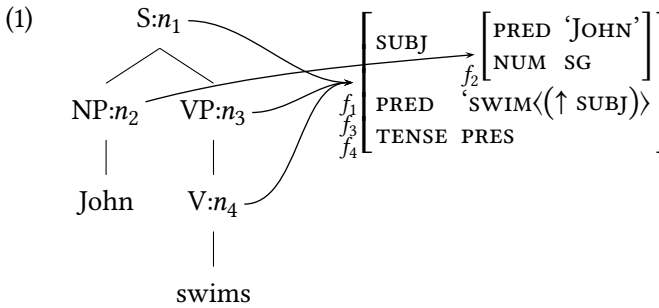
“An f-structure unit  $f$  is  $\phi$ -accessible from a node  $n$  iff either  $n$  is  $\phi$ -linked to  $f$  (that is,  $f = \phi(n)$ ) or  $f$  is contained within  $\phi(n)$  (that is, there is a chain of attributes that leads from  $\phi(n)$  to  $f$ ).” (Bod & Kaplan 1998: 146)

As an example, consider (1):

---

handled by a context-sensitive grammar, whereas English is arguably context-free. Note that Futrell et al. (2016) claim the Amazonian language Pirahã to be finite-state, so the Chomsky Hierarchy no longer seems to be particularly helpful as a characterisation of human languages in general. Nonetheless, the fact that LFG is beyond context-free would allow it to claim that it is a general enough model to cope with languages like Dutch. Note too that a grammar formalism should be sufficiently constrained to ensure that parsing can be done in polynomial time.

## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

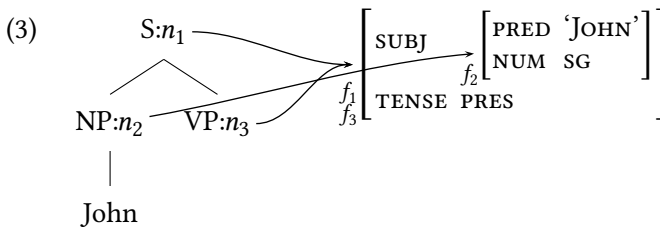


The  $\phi$ -links are shown in (2):

$$(2) \quad \phi(n_1) = f_1, \phi(n_2) = f_2, \phi(n_3) = f_3, \phi(n_4) = f_4, \phi(n_1) = \phi(n_3) = \phi(n_4)$$

$\phi$ -accessibility reflects the intuitive notion that nodes in a tree carry information only about the f-structure elements to which the root node of the tree permits access, as in (1). Note that all f-structure units are  $\phi$ -accessible from the S, VP and V nodes, but TENSE and the top-level PRED (the main verb *swim*) cannot be accessed via  $\phi$  from the subject NP node.

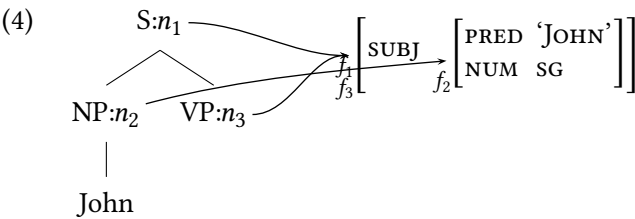
*Frontier* operates as in Tree-DOP, deleting all subtrees of the selected frontier nodes. It also deletes all  $\phi$ -links of these deleted nodes together with any semantic form (e.g. in (1), 'SWIM<((↑ SUBJ))>') as is the case if the V:swims node is deleted in (3):



This illustrates the ability of *Root* nodes to access certain f-structure features even after subnodes have been deleted. Even though the V:swims node is deleted in the c-structure tree, only the semantic form 'swim<((↑ SUBJ))>' is deleted from the f-structure, and the TENSE feature remains.<sup>4</sup>

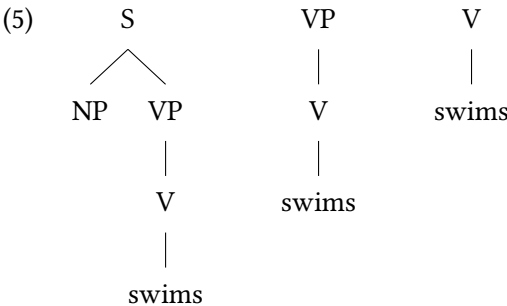
It is, however, possible to prune (3) still further, as (4) illustrates:

<sup>4</sup>Note that subject-tense agreement is seen in some languages e.g. Hindi. Accordingly, there is no universal principle which should rule out fragments such as (3).



This is achieved by applying a third, and new operation, *Discard*, to the TENSE feature in (3).<sup>5</sup> The *Discard* operation adds considerably to LFG’s robustness by providing generalized fragments from those derived via *Root* and *Frontier* by freely deleting any combination of attribute-value pairs from an f-structure except those that are  $\phi$ -linked to some remaining c-structure node, or that are governed by the local predicate (i.e. required to be present). Its introduction also necessitates a new definition of the grammaticality of a sentence *with respect to a corpus*, namely any sentence having at least one derivation whose fragments are produced only by *Root* and *Frontier* and not by *Discard*. Way (1999) splits fragments into separate bags of *Discard* and non-*Discard* fragments in order “to facilitate the consideration of grammaticality.” Bod (2000) demonstrates that this is helpful for LFG-DOP, too, on experiments with the Verbmobil and Homecentre corpora, which compare favourably with the original model of Bod & Kaplan (1998). In contrast, Hearne & Sima’an (2004) present an improved back-off estimation method where non-*Discard* fragments are naturally preferred.

We omit here the complete LFG-DOP treebank (ignoring the effects of the *Discard* operator) for the sentence *John swims*, but refer the interested reader to Figure 4.1 in Way (2001: 114). Nonetheless, as he does, we point out that each c-structure fragment in an LFG-DOP corpus is not necessarily linked to a unique f-structure fragment. From his Figure 4.1, consider the three fragments in (5):



<sup>5</sup>This function generates appropriate fragments for English which have no subject-tense dependency; accordingly, we would expect more fragments like (4) in English treebanks, but fewer such fragments for Hindi, say, given the point made in fn. 4.



## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

These three c-structure fragments all map to the same f-structure fragment in (6) because of equations such as  $\phi(n_1) = \phi(n_3) = \phi(n_4)$  in (2):

$$(6) \quad \left[ \begin{array}{ll} \text{SUBJ} & [\text{NUM SG}] \\ \text{PRED} & \text{'SWIM'}\langle(\uparrow \text{SUBJ})\rangle' \\ \text{TENSE} & \text{PRES} \end{array} \right]$$

This f-structure shows that *swims* being singular requires a singular subject. Of course, to be completely accurate, we should add in a SUBJ:PERS:3 constraint too, to prevent strings such as *I swims* and *you swims* from being deemed grammatical.

We can illustrate the effect of *Discard* in relaxing the SUBJ:NUM:SG constraint with *swims* in (7):

$$(7) \quad \begin{array}{ccc} \text{VP} & \left[ \begin{array}{ll} \text{SUBJ} & [] \\ \text{PRED} & \text{'SWIM'}\langle(\uparrow \text{SUBJ})\rangle' \\ \text{TENSE} & \text{PRES} \end{array} \right] & \text{VP} \\ | & \nearrow & | \\ \text{V} & & \text{V} \\ | & & | \\ \text{swims} & & \text{swims} \end{array}$$

Accordingly, if the ill-formed string *The men swims* were input, it could be processed by LFG-DOP because of generalised fragments like these, but would be ruled out as ungrammatical in LFG, given f-structures like (6). Note that *Discard* has been applied to the rightmost f-structure in (7).

### 2.2.3 Composition in LFG-DOP

**Composition** in LFG-DOP is also a two-step operation. C-structures are combined by leftmost substitution, as in Tree-DOP, subject to the matching of their nodes. F-structures corresponding to these nodes are then recursively unified, and the resulting f-structures are subjected to the grammaticality checks of LFG.

### 2.2.4 Probability Models for LFG-DOP

$CP(f \mid CS)$  denotes the probability of choosing a fragment  $f$  from a competition set  $CS$  of competing fragments. In Tree-DOP, we wanted to select a tree  $t$  from a treebank, whereas in LFG-DOP we are interested in selecting a  $\langle c, f \rangle$  pair from a corpus. The probability of an LFG-DOP derivation is the same as in Tree-DOP; it

*Aoife Cahill & Andy Way*

is just the derivation itself which changes. As in DOP, then, an LFG-DOP derivation  $D = \langle f_1, f_2 \dots f_n \rangle$  is produced by a stochastic branching process which at each stage in the process randomly samples from a competition set  $CS$  of competing samples, as in (8) (cf. example (10) in Bod & Kaplan 1998: 148):

$$(8) \quad P(\langle f_1, f_2 \dots f_n \rangle) = \prod_{i=1}^n CP(f_i \mid CS_i)$$

This competition probability  $CP(f \mid CS)$  is expressed in terms of fragment probabilities  $P(f)$  in (9) (cf. example (11) in Bod & Kaplan 1998: 148):

$$(9) \quad CP(f \mid CS) = \frac{P(f)}{\sum_{f' \in CS} P(f')}$$

Taking (8) and (9) together, the probability of a derivation  $f$  is calculated by multiplying together the probabilities of the fragments  $f_i$  which are composed together to form that fragment; this is analogous to how derivations are computed in Tree-DOP: there, we just have tree fragments, whereas in LFG-DOP, we have tree fragments together with their associated f-structure fragments.

In Tree-DOP, apart from the *Root* and *Frontier* operations, there are no other well-formedness checks. LFG, however, has a number of grammaticality conditions, some of which – the Completeness check at least – cannot be evaluated during the stochastic process. Accordingly, probabilities for valid representations can only be defined by sampling *post hoc* from the set of representations which are output from the stochastic process. The probability of sampling a valid representation is (10) (cf. example (12) in Bod & Kaplan 1998: 148):

$$(10) \quad P(R \mid R \text{ is valid}) = \frac{P(R)}{\sum_{R' \text{ is valid}} P(R')}$$

Bod & Kaplan (1998) note that (10) assigns probabilities to valid representations whether or not the stochastic process guarantees validity. The valid representations for a particular utterance  $u$  are obtained by a further sampling step, with their probabilities given by (11) (cf. example (13) in Bod & Kaplan 1998: 148):

5 *Treebank-driven Parsing, Translation and Grammar Induction using LFG*

$$(11) \quad P(R \mid R \text{ is valid and yields } u) = \frac{P(R)}{\sum_{R' \text{ is valid and yields } u} P(R')}$$

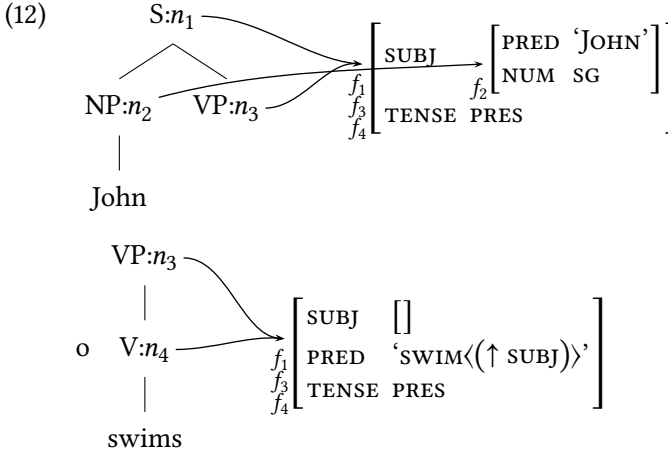
Comparing (10–11) with the equivalent formula for calculating the probability of a particular analysis for a Tree-DOP representation, Way (2001) observes that the LFG-DOP formulae contain references to *valid* structures. In Tree-DOP, apart from the root-matching criterion, there are no other validity conditions; in LFG-DOP, depending on the competition set chosen, there may be several.

Omitting the details for reasons of space, Bod & Kaplan (1998) give three different competition sets depending on the stage at which the LFG grammaticality checks are carried out, which affect the **probability models** for LFG-DOP:

1. A straightforward extension of the Tree-DOP probability model, where the choice of a fragment depends only on its *Root* node (i.e. c-structure matching category) and not on the Uniqueness, Completeness or Coherence conditions of LFG, which are enforced off-line.
2. *Root* nodes must match, and f-structures must be unifiable if two LFG fragments are to be combined. This model takes the LFG Uniqueness condition (namely that each attribute has only one value) as well as the *Root* category into account. As the resultant fragments produced vary depending on the derivation followed, unifiability must be determined at each step in the process.
3. In addition to the previous two steps, the LFG Coherence check is enforced at each step, ensuring that each grammatical function (SUBJ, OBJ etc.) present in the f-structure is governed by a PRED. This means that in this model, we are dealing only with well-formed c-structures which correspond to coherent and consistent f-structures, i.e. structures which satisfy LFG's Uniqueness check, thereby permitting unification only where exactly appropriate. As we have noted already, the LFG Completeness check can only be enforced after all other validity sampling has taken place.

Let us now return to the sentence *John swims*, and show one possible derivation of the  $\langle c, f \rangle$  pair in (1). A straightforward way of doing this would be to compose (via the 'o' operator in (12)) the  $\langle c, f \rangle$  fragment in (3) with the leftmost fragment in (7), which we include in full in (12):

Aoife Cahill & Andy Way



This is possible given that the VP node in the upper tree is vacant, so the lower VP tree can be substituted for this node. The respective  $\langle c, f \rangle$  fragment in (1). Throughout the derivation of this  $\langle c, f \rangle$  pair, we have satisfied DOP's *Root* condition (leftmost substitution of 'like' categories only), as well as the Uniqueness, Completeness and Coherence grammaticality conditions of LFG. As a consequence, the resultant structures in (1) are valid. This is equivalent to using the third option given above for possible competition sets.

Of course there will be many other possible derivations which contribute to the overall probability of the sentence *John swims*. Note that if we enforce LFG's grammaticality checks on-line, leftmost substitution of non-*Discard* fragments reduces the size of the competition set for future iterations of the composition process. In (12), for instance, enforcing the Uniqueness condition on-line (models 2 or 3 above) prevents any fragment other than a singular intransitive VP from being substituted into the VP slot. In Tree-DOP, *any* VP could be substituted at this node.

### 3 LFG-DOT

In this section, we demonstrate that problems with the LFG-MT (Kaplan et al. 1989) and Data-Oriented Translation (DOT: Poutsma (2000)) models of translation can be solved by LFG-DOT.<sup>6</sup> As the LFG-DOT models proposed by Way

<sup>6</sup>Hearne (2005) demonstrates that reasonably large-scale models can be built with DOT that considerably outperform SMT. Bod (2007) contains results which demonstrate similar improve-

## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

(1999) and Way (2001) are based on LFG-DOP, they have the same advantages as shown in the previous section, albeit now for translation:

1. Being a probabilistic model, LFG-DOT can overcome problems encountered by LFG-MT which is based solely on LFG's constraints; and
2. By appealing to LFG's f-structure constraints, LFG-DOT can overcome problems encountered by DOT which is based solely on trees.

### 3.1 LFG-MT

A translation model based on LFG was first presented in Kaplan et al. (1989). This original model introduces the  $\tau$ -correspondence as a mapping between source and target f-structures. For *swim*, we would need a transfer lexicon entry such as (13) for translation between English and French:

- (13) *swim*:  
 $(\tau \uparrow \text{PRED}) = \text{nager}$   
 $(\tau \uparrow \text{SUBJ}) = \tau(\uparrow \text{SUBJ})$

Being a straightforward translation example, this entry demonstrates two things: (i) that the translation of the verb *swim* is *nager*, and (ii) that the translation of the subject of *swim* is the subject of *nager*.

This model is very elegant, and allows for some difficult translation problems to be handled by the LFG-MT formalism. For example, verbs with different semantic forms can be handled relatively straightforwardly. Assume the translation case in (14):

- (14) The student answers the question  $\longleftrightarrow$  L'étudiant répond à la question.

This case can be dealt with as in (15):

- (15) *answer*:  
 $(\tau \uparrow \text{PRED}) = \text{répondre}$   
 $(\tau \uparrow \text{SUBJ}) = \tau(\uparrow \text{SUBJ})$   
 $(\tau \uparrow \text{OBL OBJ}) = \tau(\uparrow \text{OBJ})$

---

ments over SMT, but for *really* large-scale models at the time. Given the massive time and space constraints involved in processing DOP models, it is noteworthy that Bod was able to build DOT models trained on more than 750K sentence-pairs of German-English Europarl data (Koehn 2005).

Aoife Cahill & Andy Way

This states that *répondre* is the corresponding French predicate of *answer*, that the translation of the SUBJ is straightforward, and that the translation of the OBJ of *answer* is the OBL OBJ of *répondre*.

The LFG-MT model of Kaplan et al. (1989) can also deal correctly with the *like-plaire* relation-changing case, as (16) demonstrates:

- (16) *like*:  
 $(\tau \uparrow \text{PRED}) = \text{plaire}$   
 $(\tau \uparrow \text{OBL}) = \tau(\uparrow \text{SUBJ})$   
 $(\tau \uparrow \text{SUBJ}) = \tau(\uparrow \text{OBJ})$

That is, the subject of *like* is translated as the oblique argument of *plaire*, while the object of *like* is translated as the subject of *plaire*.

However, a line of work showed that while the  $\tau$ -equations of Kaplan et al. (1989) are by and large able to link exactly those source–target elements which are translations of each other, there are a number of cases where this machinery is unable to cope with a set of translation cases, in particular embedded headswitching examples and the correct translation of adjuncts (cf. Arnold et al. 1990; Sadler & Thompson 1991; Way 2001).<sup>7</sup>

It is worth noting that an updated version of LFG-MT was described in Kaplan & Wedekind (1993) which used the concept of *Restriction* to try to overcome some of the problems in mapping between flat syntactic f-structures to hierarchical semantic ones. However, as well as receiving criticism from a monolingual perspective (cf. Butt (1994) and complex predicates in Urdu), Way (2001) demonstrates this new approach failed to ensure that only the correct translations ensued; rather, it was left to a human expert to select the correct translation from a set of alternatives, many of which were incorrect. Despite being an improvement on the original model of Kaplan et al. (1989), it is still open to criticism as a general model of translation.

Another solution proposed around this time involved using linear logic (van Genabith et al. 1998), but this involved adding massive redundancy in the transfer lexicon, cf. Way (2001: 92–96).

---

<sup>7</sup>To give the reader some insight into the first-mentioned issue without having to consult the primary literature, LFG-MT can cope with ‘straightforward’ headswitching examples like *The baby just fell*  $\longleftrightarrow$  *Le bébé vient de tomber*. However, when such examples appear in embedded clauses, as in *I think that the baby just fell*  $\longleftrightarrow$  *Je pense que le bébé vient de tomber*, *ad hoc* solutions are required to avoid target f-structures being doubly rooted, i.e. two  $\tau$ -equations result in inconsistent solutions where one piece of f-structure is required to simultaneously fill two different slots.

## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

Note too that work continued on using LFG as a basis for MT after LFG-DOT was introduced. One such model was that of Riezler & Maxwell (2006). Note that their paper is not a comparison of LFG-MT, but rather with SMT (Koehn et al. 2003). Note that they add a ‘fragment grammar’ which “allows sentences that are outside the scope of the standard grammar to be parsed as well-formed chunks” (p.251), but they do not compare this with the bag of *Discard*-generated fragment-pairs in LFG-DOT. This work is extended by Graham & van Genabith (2012), who incorporate a deep syntax language model directly into the decoder, as opposed to using it *post hoc* to improve the grammaticality of the target translations. Note that neither approach shows how their models handle any of the traditional ‘hard’ translation cases. For the approach of Riezler & Maxwell (2006), being based on transfer rules – albeit automatically extracted ones – it will surely fail in similar ways to LFG-MT. As to the model of Graham & van Genabith (2012), and approaches based on SMT in general, it is doubtful whether the system designers can answer the question as to how such translational phenomena is handled, as SMT does not work in this way. Of course, test sets can be designed which include such ‘hard’ cases, and the translation output inspected, but SMT systems by their very nature are far less inspectable than systems which include syntactic constraints, so even if such sentences were translated correctly, it would be hard to know why exactly. Of course, this problem is worse again with today’s state-of-the-art neural models; despite the improved quality that can be derived, our knowledge as to what is going on internally inside the systems is less than it’s ever been!

### 3.2 Data-Oriented Translation

Poutsma (2000) produced two models of tree-based translation, DOT1 and DOT2. These models were formulated along the same lines as DOP and LFG-DOP, with definitions of the representations to be used, how these were to be decomposed, recomposed, and a probability model.

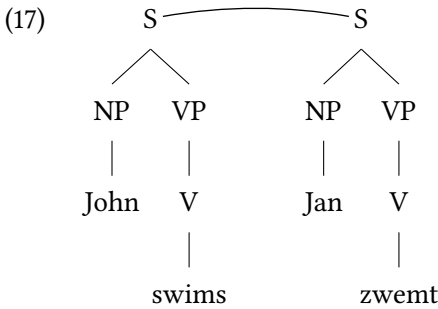
In DOT, the latter determined the likelihood of a target translation given a source string. The representations used were PS trees, decomposition described how to extract well-formed subtree-pairs from these representations,<sup>8</sup> and the composition operator used was leftmost substitution (to ensure unique derivations) of matching *Root* labels.

---

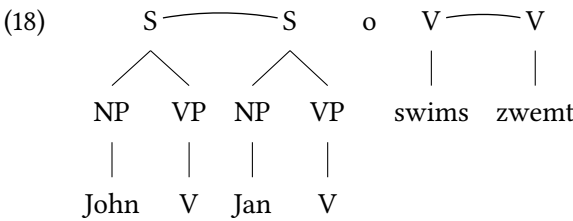
<sup>8</sup>In his thesis, Way 2001 introduces the label  $\gamma$  to refer to the function that links DOT source and target subtree fragments. See Section 3.3.2 for models which use the  $\gamma$  function in LFG-DOT, and Poutsma 2000: Sect. 2.1 for a description of how linked subtrees like the V-labelled fragments in (18) are extracted from tree pairs such as (17).

Aoife Cahill & Andy Way

We illustrate a linked translation pair in DOT in (17) for the the sentence pair  $\langle \textit{John swims}, \textit{Jan zwemt} \rangle$ :<sup>9</sup>



If we assume that the sentential fragment in (17) is unseen in our DOT treebank, one derivation of the translation *Jan zwemt* given the source sentence *John swims* might be (18):



Way (1999) showed that the DOT1 model could not always explicitly relate parts of the source-language structure to the corresponding, correct parts in the target structure, so fails to translate correctly where source and target strings differ with respect to word order (e.g. the *like*  $\Leftrightarrow$  *plaire* relation changing case – which LFG-MT can handle, cf. (16) – plus many more ‘hard’ translation cases described in Way et al. (1997)).

DOT2 was developed as a consequence of these failings, and improves over DOT1 by not restricting the composition operation to left-most substitution on *both* sides. With that change, DOT2 manages to overcome cases of word-order difference by and large. However, Way (2001) notes that:

“this is compromised by a lesser amount of compositionality in the translation process. Given the small number of fragments playing a role in the

<sup>9</sup>Here we ‘translate’ names to indicate that the translation process has been successful, as opposed to merely passing over a source word as untranslated – an out-of-vocabulary item – into the target side.



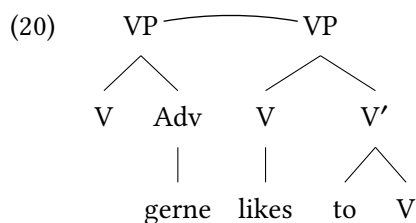
## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

derivation of some translations involving complex phenomena, almost the exact linked sentence pair may need to be present in order for a translation to be possible. Furthermore, any such translations produced have extremely small probabilities with respect to the corpus. Finally, of course, translation systems which are based purely on PS trees will ultimately not be able to handle certain linguistic phenomena.” (Way 2001: 190)

To illustrate the ‘limited compositionality’ problem in DOT2, Way (2001) appeals to the translation pair in (19).<sup>10</sup>

- (19) DE: Johannes schwimmt gerne ⇔ EN: John likes to swim.

Essentially, the VPs cannot be broken down further; *schwimmt* and *swim* are not translationally equivalent – one is inflected and the other is in the infinitive form – so in their source–target tree pairs, links cannot be drawn between the fragment-pair in (20), as we might otherwise wish to do, in order to describe the basic translation relations in (19):



Accordingly, while it is possible for DOT2 to cope with such examples in contrast to DOT1, which couldn’t handle them at all, the exact VPs (*likes to swim*, here) have to exist *a priori* in the treebank. This is because these linked VP pairs are handled non-compositionally in DOT2 between German and English, but the monolingual VPs are treated compositionally in DOP. As can be seen, DOT2 approximates to a translation dictionary for such cases – as *likes to* can be followed by pretty much any verb in English, and *gerne* can modify pretty much any verb in German – which is clearly impractical, and so can be disregarded as a general model of translation.

<sup>10</sup>Given that other similar cases exist, e.g. DE: Josef läuft zufällig ⇔ EN: Joseph happens to run, the redundancy in the DOT2 approach really shows itself to be problematic when such cases are combined, as in strings such as *John likes to happen to swim* (i.e. John likes to swim by chance, rather than planning ahead), and *John happens to like to swim*.

*Aoife Cahill & Andy Way*

### 3.3 Combining DOT and LFG-MT: the Best of Both Worlds

In his thesis, Way (2001) provides four LFG-DOT models which solve all these ‘hard’ cases:

1. Model 1: Translation via  $\tau$
2. Model 2: Translation via  $\tau$  and  $\gamma$
3. Model 3: Translation via  $\gamma$  with Monolingual Filtering
4. Translation via  $\gamma$  and ‘Extended Transfer’

#### 3.3.1 LFG-DOT1

Way (2001) describes this as a simple linear model, as in (21):

$$(21) \quad \begin{array}{ccc} & \text{LFG-DOP-}\phi & \\ c & \text{-----} & f \\ & & | \tau \\ c' & \text{-----} & f' \\ & \text{LFG-DOP-}\phi' & \end{array}$$

The different components needed are:

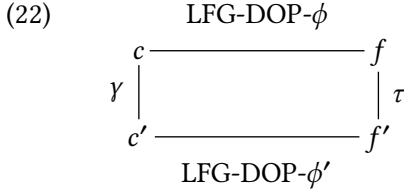
- a source language LFG-DOP model;
- the  $\tau$  mapping;
- a target language LFG-DOP model.

Way (2001: 193) notes that “LFG-DOT1 contains two monolingual LFG-DOP language models ... [so] *Discard* can be run on both source and target sides. This means that LFG-DOT1 can cope with ill-formed or previously uncovered input which LFG-MT would not be able to handle at all”. Despite this advantage, LFG-DOT1 unsurprisingly suffers from the same problems as LFG-MT, as its translation function is described by the same operator  $\tau$ .

## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

### 3.3.2 LFG-DOT2

Given that  $\tau$  is an insufficient operator to define all translation problems (cf. fn. 7, for example), Way (2001) describes the translation relation using both the  $\gamma$  and  $\tau$  functions in his LFG-DOT2 model, summarised in (22):



This is clearly a more complex model than LFG-DOT1, necessitating:

- a source language LFG-DOP model;
- the  $\gamma$  mapping (i.e. the DOT2 model of translation, cf. fn. 8);
- a target language LFG-DOP model;
- a probabilistic transfer component.

Way (2001) provides a number of ways in which the  $\gamma$  and  $\tau$  functions might co-operate in his LFG-DOT2 model. He notes that using LFG-DOP as the source and target language models overcomes the shortcomings of both Tree-DOP and LFG, and that including  $\tau$  allows certain ‘hard’ cases (like relation-changing) to be handled correctly, unlike the DOT1 model.

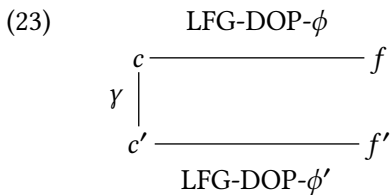
Furthermore, Way (2001) notes that LFG-DOT2 is more robust than LFG-MT, in that *Discard* can produce generalized fragments which may be able to deal with input for which LFG-MT cannot offer any translation.

Ultimately, as the  $\tau$  mapping cannot always produce the desired translation, Way jettisons this function in his LFG-DOT3 and LFG-DOT4 models, to which we turn next.

### 3.3.3 LFG-DOT3

LFG-DOT3 relies solely on  $\gamma$  to express the translation relation. The architecture of LFG-DOT3 is shown in (23):

Aoife Cahill & Andy Way



Way (2001) demonstrates that contrary to other models described here, embedded headswitching cases in LFG-DOT3 are handled in the same manner as non-embedded headswitching cases, exactly as required (cf. fn. 7). He also shows that LFG-DOT3 can cope with certain cases of combinations of exceptional phenomena which prove problematic for other formalisms. However, like DOT2 (cf. Section 3.2), LFG-DOT3 also suffers from the problem of limited compositionality.

### 3.3.4 LFG-DOT4

To overcome this problem, Way (2001) uses a restricted form of *Discard* in an ‘extended transfer’ phase in LFG-DOT4 to generalize the translation relation appropriately. Essentially, in LFG-DOP (and consequently LFG-DOT), fragments generated by *Discard* occupy an unjustifiably large proportion of the probability space. Accordingly, Way (1999) proposes to split fragments into two bags: those generated by *Root* and *Frontier*, and those generated by *Discard*. In LFG-DOT4, Way (2001) allocates a small amount of the probability space to lemmatized translation pairs produced by a second application of *Discard*.<sup>11</sup> To revisit the problematic example in (19), if *Discard* is used to relax the TENSE constraint, then the V nodes in (20) can be linked; they couldn’t before as the V in German was a finite verb, while the V in English was an infinitive. Accordingly, Way (2001: 190) suggests that “this model describes the translation relation exactly as required, and furthermore overcomes the problems of LFG-MT ... and DOT models of translation”. See Table 1 for a summary of the comparative advantages and disadvantages of each of the models covered in this chapter.<sup>12</sup>

<sup>11</sup>Another way of mitigating this problem is suggested by Way (2001: 112), namely to adopt the approach of Zaenen & Kaplan (1995), which cuts down on the possible number of LFG-DOP fragments compared to the description of LFG in Kaplan & Bresnan (1982). In Zaenen & Kaplan (1995), lexical heads are  $\phi$ -linked only to semantic forms and not to their enclosing f-structures, while other primitive feature values remain unlinked.

<sup>12</sup>See Way (2003) for more details on these models, and Hearne (2005) for an alternative LFG-DOT model based on LFG-DOT3 but which incorporates a different probability model and fragmentation procedure.

## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

Table 1: A comparison of the advantages and disadvantages of the MT models described in this work

Model	Ill-formed Input	Word Order	Embedded Headswitching	All ‘hard’ Cases	Avoids Limited Compositionality
LFG-MT	N	Y	N	N	N
DOT1	Y	N	N	N	N
DOT2	Y	Y	Y	N	N
LFG-DOT1	Y	Y	N	N	N
LFG-DOT2	Y	Y	Y	N	N
LFG-DOT3	Y	Y	Y	Y	N
LFG-DOT4	Y	Y	Y	Y	Y

## 4 Automatic Derivation of F-structures from Treebanks

In this section we consider how the resources needed for large-scale LFG-DOP and LFG-DOT models can be generated. We also explain the two different ways in which f-structures can be derived from a tree.

### 4.1 Towards large-scale resources for LFG-DOP and LFG-DOT

LFG-DOP needs large collections of monolingual annotated data (treebanks) in order to parse monolingual input, and LFG-DOT needs large collections of bilingual annotated data. At the time LFG-DOP and LFG-DOT were being developed, no such large f-structure annotated data existed. Constituency treebanks had been available for several years, and large-scale hand-crafted LFG grammars were available only for a few languages. However, neither could provide the input needed to support the training of LFG-DOP or LFG-DOT models. Constituency treebanks alone could not provide the linguistic detail needed, and hand-crafted grammars were unable to select the most likely parse from a (sometimes) large number of possible solutions.

To address these shortcomings, van Genabith, Way, et al. (1999b) and van Genabith, Sadler, et al. (1999) proposed initial methods to automatically derive the LFG-treebank resources required to support training LFG-DOP and LFG-DOT models, although this was not the main driving force behind this work.

Initially, the work conducted produced grammars and lexicons for English, which seeded high-performing probabilistic parsers (see Section 5). Later, related

*Aoife Cahill & Andy Way*

methods were used to extract similar resources for a range of other languages (cf. Section 6). Once the general approach had been validated for different languages and treebanks, it is possible to sketch a research project which could generate the resources needed for large-scale LFG-DOP and LFG-DOT experimentation.

Taking a large-scale parallel corpus such as Europarl (Koehn 2005), we would need to:

1. Parse source and target sides to generate c-structure trees for the two languages;
2. Run the f-structure annotation algorithm(s) over each side;
3. Apply the *Root* and *Frontier* operations to extract the separate bags of fragments.

After step 2, we have  $\langle c, f \rangle$  pairs of structure for all sentences on both sides of the corpus, so we can build LFG-DOP models for the individual source and target sides by running *Root* and *Frontier* operations on each side, and start producing  $\langle c, f \rangle$  pairs for new monolingual input. To generate resources for the better of the four models, LFG-DOT4, we need to align each source tree generated in step 1 with each target tree generated in the same step. Fortunately, Europarl contains information regarding which sentences in one language map to which sentences in another, so we can now apply *Root* and *Frontier* operations on both sides to extract the separate bags of fragments that are needed, and start translating new input strings. This experiment remains for future work.

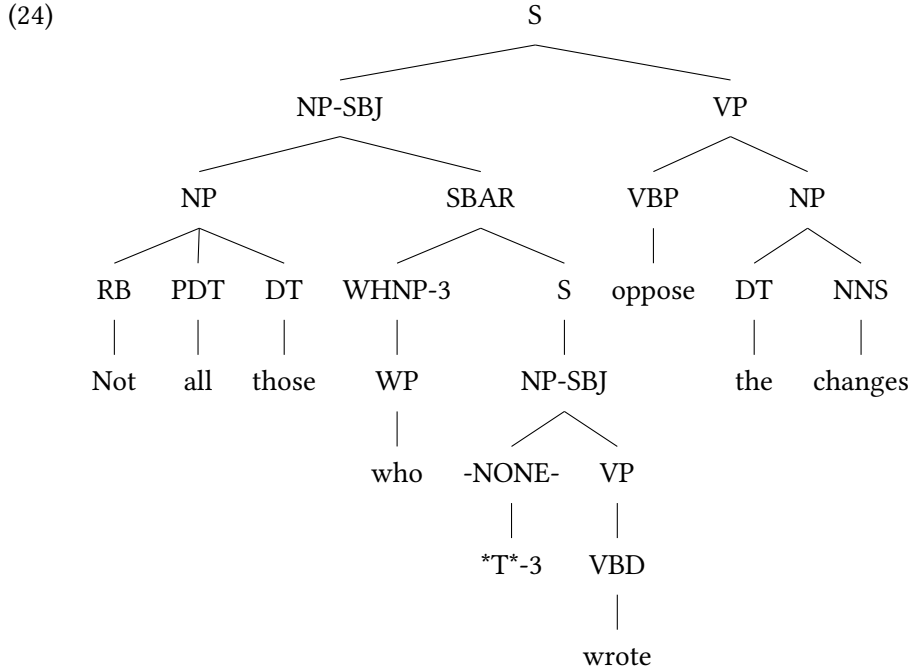
## 4.2 Direct Transformation vs. Indirect Annotation

The initial approaches of van Genabith, Way, et al. (1999b) and van Genabith, Sadler, et al. (1999) focused on deriving f-structure annotations from PS trees. The intuition was that there were already reasonably reliable tools for automatically producing a tree from an input sentence, and so it would be easier to scale a tree-annotation plus f-structure derivation approach, compared to automatically deriving c- and f-structure simultaneously from raw input.

There are two ways to derive an f-structure from a tree: **direct** transformation or **indirect** annotation. The direct method recursively and destructively transforms a treebank tree into an f-structure. The indirect method only ever adds information: it annotates the treebank tree with f-structure annotations (equations). These annotations are then collected and passed to a constraint solver which resolves the equations and, if the equations are consistent, outputs an f-structure.

## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

Examples (24)–(26) illustrate the indirect method: all nodes in the tree in (24) are annotated with equations in (25), which are collected and resolved into an f-structure in (26).



(25) (S  
 (NP-SBJ[up-subj=down]  
 (NP[up=down]  
 (RB[down-elem=up:adjunct] Not[up-pred='not'])  
 (PDT[up-spec:det=down] all[up-pred='all'])  
 (DT[up=down] those[up-pred='those'])  
 )  
 (SBAR[up-relmod=down]  
 (WHNP-3[up-topicrel=down,up-topicrel:index=3]  
 (WP[up=down] who[up-pred=pro,up-pron\_form='who'])  
 )  
 (S[up=down]  
 (NP-SBJ[up-subj=down,up-subj=up:topicrel]  
 (-NONE- \*T\*-3)  
 )  
 (VP[up=down]  
 (VBD[up=down] wrote[up-pred='write',up-tense=past])  
 )  
 )  
 )  
 )

*Aoife Cahill & Andy Way*

```

    )
  )
)
(VP[up=down]
  (VBP[up=down] oppose[up-pred='oppose',up-tense=pres])
  (NP[up-obj=down]
    (DT[up-spec:det=down] the[up-pred='the'])
    (NNS[up=down] changes[up-pred='change', up-num=pl,up-pers=3])
  )
)
)
(. .)
)

(26) subj : adjunct : 1 : pred : not
      spec : det : pred : all
      pred : those
      relmod : topicrel : index : 3
                pred : pro
                pron_form : who
      subj : index : 3
        pred : pro
        pron_form : who
      pred : write
      tense : past

      pred : oppose
      tense : pres
      obj : spec : det : pred : the
        pred : change
        num : pl
        pers : 3

```

The earliest approach to automatically identifying functional grammatical categories such as SUBJ, OBJ, etc in PS trees is probably that of Lappin et al. (1989). Nodes in trees are linked to their corresponding grammatical functions. Their motivation was to generate a set of grammatical function-based transfer rules as part of an MT project.

A regular expression-based, indirect automatic annotation method is described in Sadler et al. (2000). This involves extracting a context-free PS grammar (CFG) from a treebank fragment. F-structure annotation principles are stated in terms of regular expressions matching CFG rules. By applying regular expression-based annotation principles to the rules that are extracted, and using these annotated rules to re-match the original trees, f-structures can be generated for these trees. The number of annotation principles is appreciably smaller than the number of



## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

extracted CFG rule types since the regular expression-based annotation principles capture linguistic generalisations.

The flat, set-based tree description rewriting method of automatically annotating trees with f-structure descriptions developed by Frank (2000) can be seen as a generalisation of the regular expression-based technique of Sadler et al. (2000). Here the idea is that each tree is translated into a flat description using terms from a tree description language (e.g. *lex*, *arc*, *phi* etc.). Annotation principles are then defined in terms of rules employing a rewriting system originally developed for transfer-based MT architectures (Kay et al. 1994). In certain circumstances, the principles can be applied order-independently, or in a particular cascading order. One of the advantages of this method is that tree fragments of arbitrary depth can be considered, whereas in the regular expression-based method, tree depth is limited to 1 (i.e. CFG rules).

The earlier approaches were limited in scale to corpora in the order of hundreds of trees. In Cahill et al. (2002a) a first version of a large-scale indirect annotation algorithm was described. This algorithm was scaled to a corpus containing tens of thousands of trees. The algorithm recursively traverses a PS tree and annotates f-structure information on each node. McCarthy (2003) and Burke (2006) continued to expand this algorithm in terms of linguistic coverage. The algorithm itself is modular and separates the linguistic data from the traversal algorithm. There are two stages to the algorithm: (i) “proto”-f-structures are generated which contain unresolved long-distance dependencies (LDDs); and (ii) trace information encoded in the treebank is used to correctly link moved constituents to where they should be interpreted semantically. Given a PS tree with f-structure-annotated nodes, a constraint solver based on the one described in Gazdar & Mellish (1989) was used to produce the final f-structure representation for the tree. This body of work yielded the first large-scale algorithm for converting a treebank into a corpus of f-structures. This was a prerequisite for the parsing work that built on this corpus as described in the Section 5.

Similar efforts to automatically acquire wide-coverage grammars for TAG (Xia 1999), HPSG (Miyao et al. 2003), and CCG (Hockenmaier & Steedman 2002) appeared around the same time as the work on LFG.

## 5 Probabilistic Parsing & Lexicon Induction Using LFG

With the availability of large-scale f-structure-annotated treebanks, it was now possible to train probabilistic LFG parsers.

The initial parsing experiments of Cahill et al. (2002b) were conducted on the Penn Treebank (Marcus et al. 1994). Two main approaches were compared:

Aoife Cahill & Andy Way

1. Parse with a standard CFG parser and then automatically annotate the resulting tree (*pipeline architecture*)
2. Automatically annotate the nodes in the trees of a large corpus with f-structure information and train a probabilistic parser on it (*integrated architecture*)

Both approaches yielded c-structures whose node labels included f-structure annotations. These f-structure annotations were then collected and resolved to generate a final f-structure. Initial parsers generated what were called “proto”-f-structures which did not include any LDD resolution. It should be noted that since these techniques were probabilistic, a set of n-best trees (and therefore f-structures) could also straightforwardly be produced. This was not possible with hand-crafted grammars which output all possible f-structure solutions for a given sentence without any way to sort them. Riezler et al. (2002) showed that it was possible to *post hoc* rank the output of such a parser, however.

In Cahill (2004) and Cahill et al. (2004), additional functionality was added to the original algorithm to allow for LDD resolution. This yielded more complete f-structures. There were two main components to the algorithm: (i) a set of possible functional uncertainty paths, and (ii) a subcategorisation lexicon.

In order to obtain the set of possible functional uncertainty paths, all observed paths between co-indexed material were extracted from the f-structures automatically derived from the Penn Treebank. These paths were associated with probabilities. O’Donovan et al. (2004) and O’Donovan (2006) describe an approach for automatically acquiring a large-scale subcategorisation lexicon from the Penn Treebank. This relies on the intuition that if the original conversion of the treebank into f-structures is of high enough quality, then the lexical entries for all predicates can be reverse-engineered (van Genabith, Way, et al. 1999a). Frames are not predefined, yet the frames that are automatically acquired fully reflect LDDs in the source data-structures, discriminate between active and passive frames, and conditional probabilities are associated with each frame.

Given a set of semantic forms  $s$  with probabilities  $P(s|l)$  (where  $l$  is a lemma), a set of paths  $p$  with  $P(p|t)$  (where  $t$  is either TOPIC, TOPICREL or FOCUS) and an f-structure  $f$ , the core of the algorithm to resolve LDDs recursively traverses  $f$  to identify the most likely location of co-indexed material.

Evaluation of the f-structures produced by both parsing approaches was carried out against several corpora over time: the DCU-105 corpus (Cahill et al. 2002a), the automatically converted Section 23 of the Penn Treebank, the PARC 700 corpus (King et al. 2003) and the CBS 500 (Carroll et al. 1998). F-structures

## 5 *Treebank-driven Parsing, Translation and Grammar Induction using LFG*

were converted into dependency triple format and compared to the gold-standard triples to give results in terms of precision, recall and f-score. Results demonstrated state-of-the-art results compared to other ‘deep’ parsers available at the time. Cahill, Burke, O’Donovan, Riezler, et al. (2008) summarize a large set of parser comparisons, and show that the f-structures produced by the automatic processes described above were able to outperform two hand-crafted parsers: RASP (Carroll & Briscoe 2004) and the the English ParGram LFG run on XLE (Riezler et al. 2002). Rimell et al. (2009) conduct a comparison of several “deep” parsers on a specialized corpus of sentences containing only LDDs. They find that the HPSG and CCG parsers perform better than the DCU LFG parser on this set of difficult sentences.

## 6 Multilingual Probabilistic LFG Induction

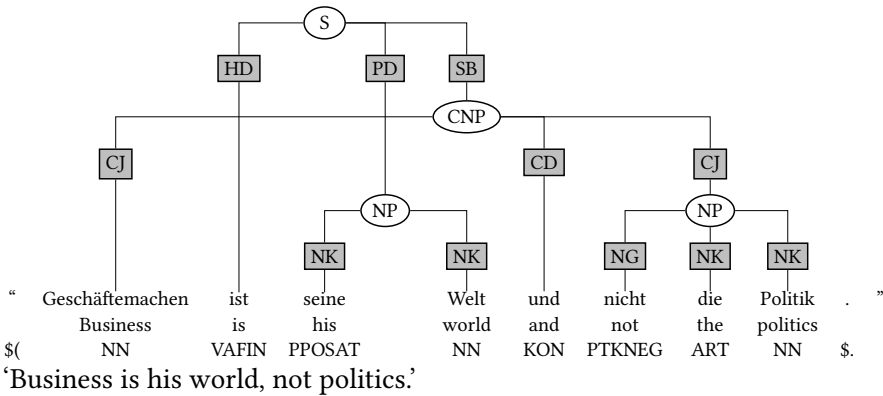
The approach developed for English was language-independent. Given a large enough and detailed treebank, one could theoretically follow the same framework to generate parsers and lexicons for other languages. Indeed, given that comparable treebank existed for some languages other than English, a large body of work ensued in this direction.

Cahill et al. (2003) first attempted this for German using the TiGer Treebank (Brants et al. 2002). This treebank differs from the Penn Treebank in that it encodes parses in terms of labeled graphs that allow crossing edges. In Cahill et al. (2003), the graphs are first converted to trees similar to those found in the Penn Treebank with trace information added to account for moved constituents. A set of rules was then developed that automatically assigned f-structure equations to the nodes in the trees, and the same techniques described in Cahill et al. (2002b) were used to automatically acquire the first large-scale probabilistic LFG for German.

We provide an example graph from the TiGer treebank in (27) for the German sentence “Geschäftemachen ist seine Welt und nicht die Politik” (“Business is his world, not politics”).

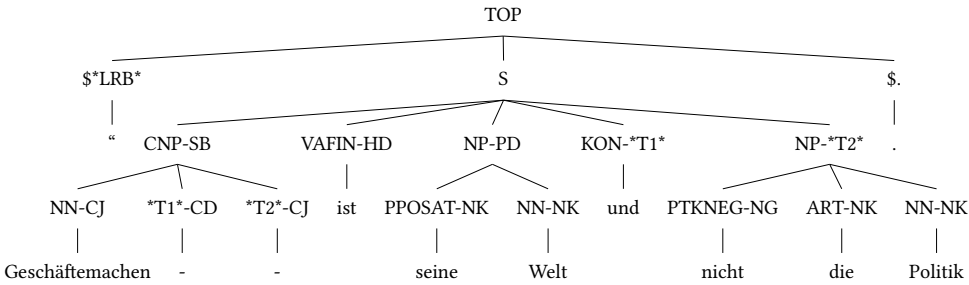
(27)

Aoife Cahill & Andy Way



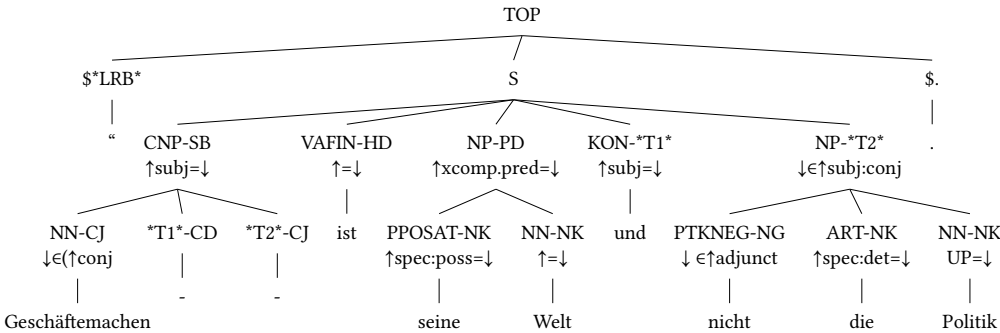
In (28), the graph in (27) is first automatically converted into a PS tree with traces and coindexation to indicate linked elements, analogous to how this kind of information is encoded in the English Penn-II treebank.

(28)



In (29), the tree in (28) is then annotated with f-structure equations. The annotation algorithm relies heavily on the functional component of the tree node labels (e.g. that SB indicates a SUBJECT).

(29)



## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

Finally, the equations in (25) are collected and passed through a constraint solver to generate the f-structure in (30), using the same procedure as for English.

$$(30) \left[ \begin{array}{cc} \text{XCOMP.PRED} & \left[ \begin{array}{c} \text{SPEC} \left[ \text{POSS} \left[ \text{PRED PRO} \right] \right] \\ \text{PRED WELT} \end{array} \right] \\ \text{SUBJ} & \left[ \begin{array}{c} \text{CONJ} \left\{ \begin{array}{c} \left[ \text{PRED GESCHÄFTEMACHEN} \right] \\ \left[ \text{PRED POLITIK} \right] \\ \left[ \text{SPEC} \left[ \text{DET} \left[ \text{PRED DIE} \right] \right] \right] \\ \left[ \text{ADJUNCT} \left\{ \left[ \text{PRED NICHT} \right] \right\} \right] \end{array} \right\} \\ \text{COORD-FORM UND} \end{array} \right] \\ \text{PRED} & \text{IST} \end{array} \right]$$

Rehbein & van Genabith (2009) continued this work and explored the effect of the design of the treebank on the success of the technique. They compared extracting a probabilistic LFG from both TiGer and TüBa-D/Z (Telljohann et al. 2006) and found (1) that automatically inducing linguistic resources from (semi-) free word order languages such as German is much harder than for more configurational languages like English, and (2) that the treebank encoding can have a significant effect on the success of the automatic f-structure annotation approach. Rehbein & van Genabith (2009) found that the encoding of linguistic structures in the TiGer treebank was better suited for automatic induction of LFG resources, because it was more difficult to automatically learn the grammatical function relations as they were encoded in the TüBa-D/Z.

For Chinese, Burke, Lam, et al. (2004) first applied the approach to the Penn Chinese Treebank (Xue et al. 2002). We provide in (31) an example tree from this treebank.

- (31) (IP-HLN  
       (NP-PN-SBJ  
           (NR 江泽民)  
           (NR 李鹏))  
       (VP  
           (VV 电唁)  
           (NP-OBJ  
               (NP-PN  
                   (NR 尼克松))  
               (NP  
                   (NN 逝世))))))  
       “江泽民李鹏电唁尼克松逝世”

*Aoife Cahill & Andy Way*

‘Jiang Zemin and Li Peng condoled the bereavement of Nixon by a telegram.’

Each node in the tree in (31) is then annotated with f-structure equations, and the f-structure in (32) is derived.

```
(32)  subj :   coord_form : null
        coord :   1 :   pred : '江泽民'
                    pers : 3
                    noun_type : proper
                    gloss : 'Jiang_Zemin'
                2 :   pred : '李鹏'
                    pers : 3
                    noun_type : proper
                    gloss : 'Li_Peng'

        pred : '电唁'
        gloss : condole_by_a_telegram
        obj :   adjunct : 3 :   pred : '尼克松'
                                pers : 3
                                noun_type : proper
                                gloss : 'Nixon'

        pred : '逝世'
        pers : 3
        noun_type : common
        gloss : 'bereavement'

        “江泽民李鹏电唁尼克松逝世”
        “Jiang Zemin and Li Peng condoled the bereavement of Nixon by a telegram.”
```

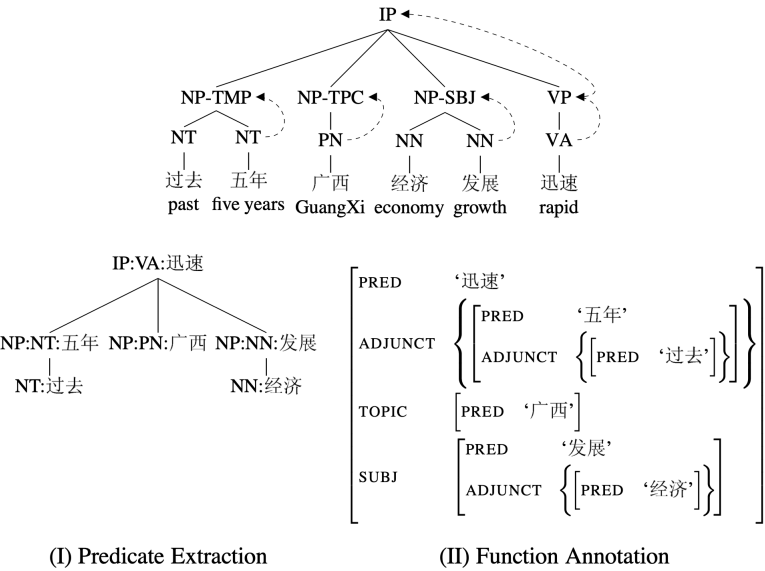
Guo et al. (2007) and Guo (2009) extended this work in terms of coverage, robustness, quality and fine-grainedness of the resulting LFG resources. They propose a more general two-stage annotation architecture, avoiding some of the limitations of the PS annotation-based method. They argue that this approach may be more suitable for less configurational languages. This algorithm works by transducing the tree into an f-structure by means of an intermediate dependency structure.

In (33), we show an example where predicate information is first extracted from the tree, and then a simpler set of function-based annotations converts the intermediate structure into an f-structure. The advantages of this approach are

5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

that it guarantees a single connected f-structure, as well as simplifying the process of taking LDDs into account.

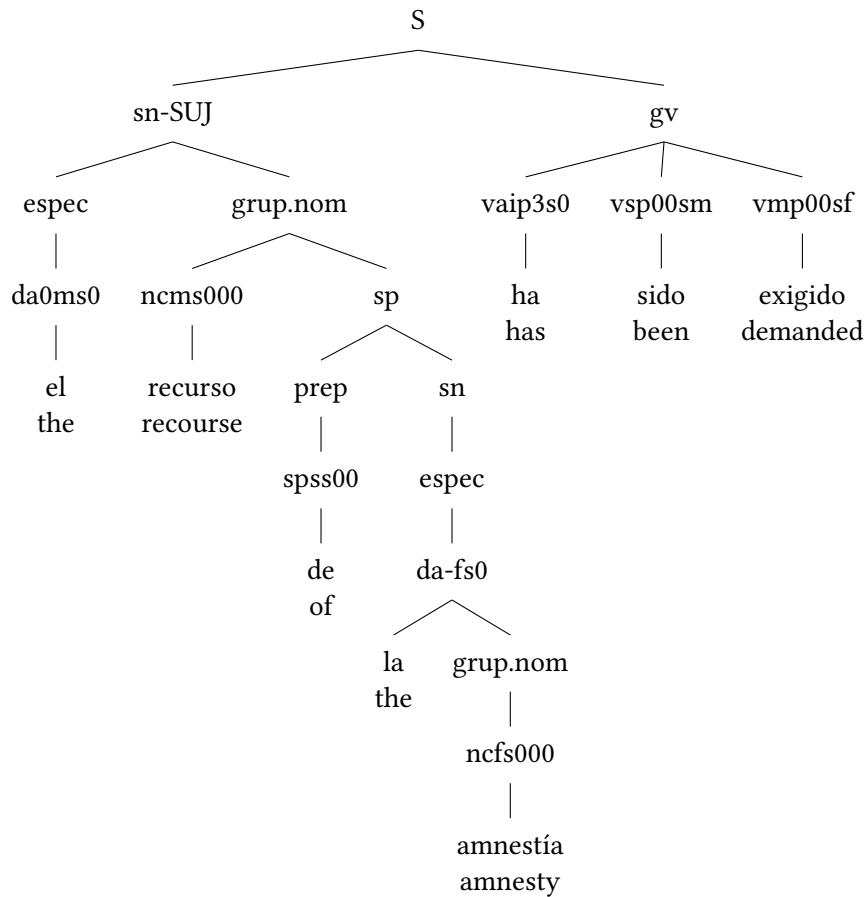
(33)



O'Donovan et al. (2005) proposed an adaptation of the original approach for Spanish using the CAST3LB treebank (Civit 2003). In (34), we provide an example from the CAST3LB treebank.

(34)

*Aoife Cahill & Andy Way*

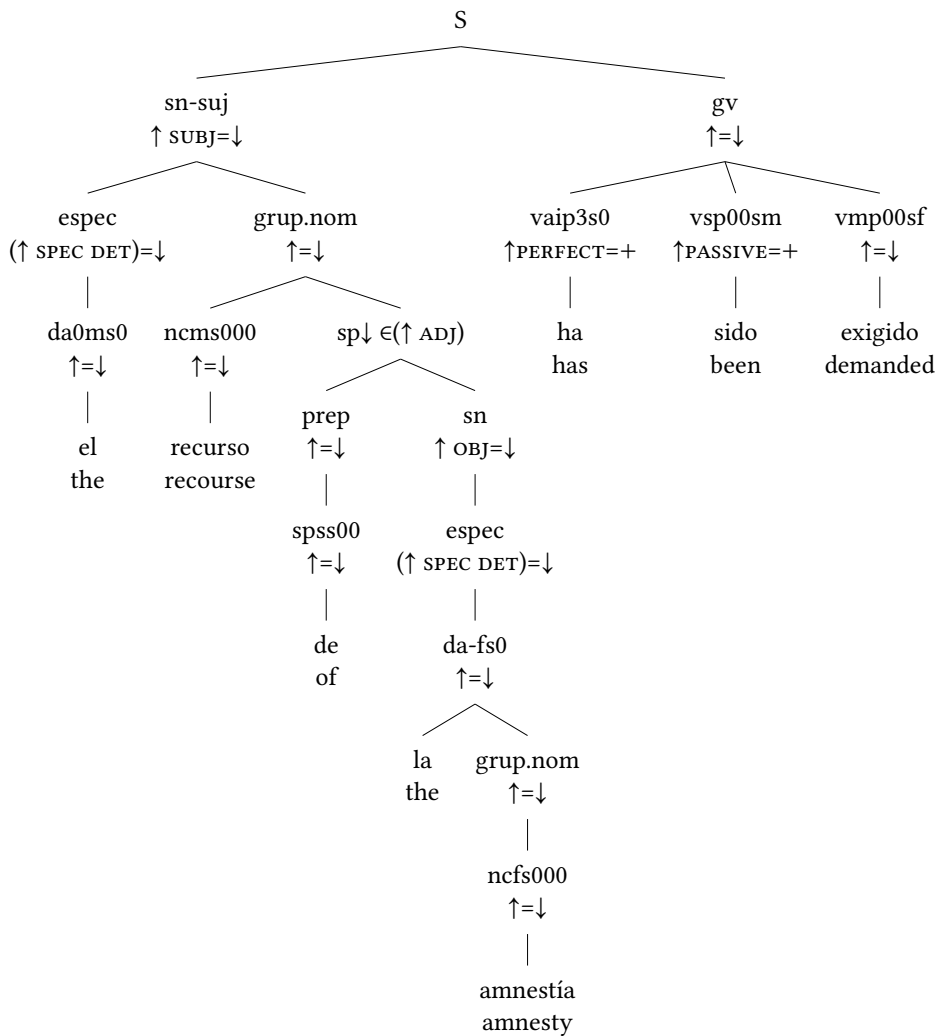


The tree in (34) is then annotated with equations, as illustrated in (35). The equations are then resolved into the f-structure in (36).

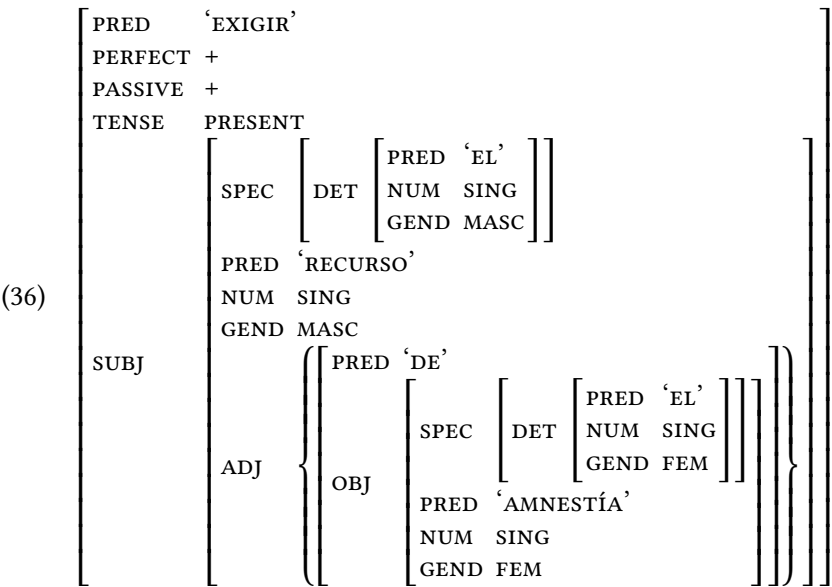
(35)



5 *Treebank-driven Parsing, Translation and Grammar Induction using LFG*



*Aoife Cahill & Andy Way*



This was extended in Chrupała & van Genabith (2006) where three main issues were addressed: (i) new constructions that had standard LFG analyses (e.g. clitic doubling and null subjects); (ii) new constructions where no LFG analysis was available (e.g. periphrastic constructions in Spanish, see Figure 1); and (iii) limitations of the previous approach due to treebank- and language-specific assumptions which did not hold for Spanish and the CAST3LB treebank. Similar to what Guo et al. (2007) and Rehbein & van Genabith (2009) had found in their adaptations, the original approach assumed that the functional information could easily be derived from the tree configuration, but this proved not to be the case for many languages. Therefore, the functional tags in the parser output were critical for the success of these annotation algorithms. As a result, Chrupała & van Genabith (2006) outlined an improved method for tagging functions in parse trees, not only for Spanish, but for English, too. This was an important step in the development of a probabilistic Spanish LFG parser based on the CAST3LB treebank.

In the case of French, no suitable treebank was immediately available. Therefore, Schluter & van Genabith (2007) first modified the Paris 7 Treebank (Abeillé et al. 2004), as this was the closest in format to what would be needed. A subset of the original treebank was transformed to yield a leaner, more coherent, treebank with several transformed structures, and new linguistic analyses. In Schluter & van Genabith (2008), it was shown that a probabilistic parser trained on the cleaner, modified treebank performed better than a parser trained on the

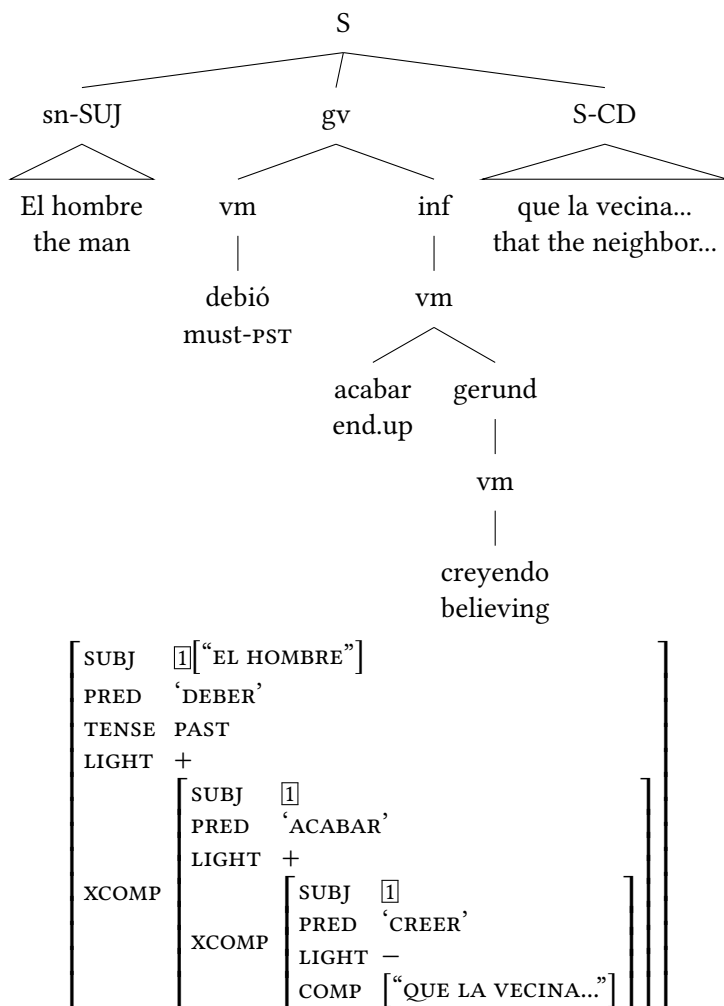
5 *Treebank-driven Parsing, Translation and Grammar Induction using LFG*

Figure 1: Treatment of periphrastic constructions outlined in Chrupala & van Genabith (2006)

*Aoife Cahill & Andy Way*

much larger, but noisier, original treebank. In addition, Schluter & van Genabith (2008) and Schluter (2011) showed that the techniques for automatically acquiring LFG resources from treebanks could successfully be adapted to the French case. Thanks to a rich morphological and functional annotation in the treebank, the automatic annotation algorithm can rely on node labels rather than inferring functional labels via tree configurations. This leads to fewer incomplete f-structures, and fewer cases where LDDs have not been resolved.

Oya & van Genabith (2007) showed that the approach can also be adapted for Japanese using the Kyoto Text Corpus (Kurohashi & Nagao 1997). The Japanese corpus encodes syntactic units in addition to rich morphological information. The automatic annotation algorithm adds f-structure equations at the level of syntactic unit. Figure 2 shows how the f-structure equations are added to each syntactic unit of the sentence “Taro went to Seoul”. In the case of Japanese LFG parsing, the key to successful parsing results was in zero pronoun identification.

Finally, Tounsi et al. (2009a) and Tounsi et al. (2009b) demonstrated that the approach was also possible for Arabic using the Penn Arabic Treebank (Maamouri & Bies 2004). The annotation algorithm was able to take advantage of rich morphological tags in the treebank to support the fact that Arabic is a morphologically rich language.

In most cases we observe that the original reliance on tree configurations to identify functional properties worked best for English. For the other languages, relying on functional information already in the original treebank, and then ensuring that the CFG parser also contains that information, yielded the most accurate f-structure parsers. Evaluation of LFG parsing for the other languages followed roughly the same procedure as for English, using a small manually annotated corpus of sentences from the treebank used to derive the algorithm and parser.

## 7 Related approaches to grammar induction

A natural evaluation of this approach to creating large-scale probabilistic LFG parsers is to compare large-scale grammars created manually using the XLE platform.

The method proposed in Riezler et al. (2002) provides a mechanism for ranking all possible solutions generated by the hand-crafted grammar, relying on the same kinds of treebank resources as the methods described above. Kaplan et al. (2004) show that the accuracy of the hand-crafted grammar is more accurate than the Collins (1999) parser (f-score of 77.6 vs 74.6), while only slightly slower

## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

#S-ID:950101001-001

\* 0 2D

太郎 たろう \* 名詞 人名 \* \* (Taro Noun Person\*\*)

が が \* 助詞 格助詞 \* \* (ga particle Case \*\*)

F0:pred='Taro',

F0:case='ga',

F2:subj=F0,

\* 1 2D

ソウル そうる \* 名詞 地名 \* \* (souru "Seoul" \* Noun Place\*\*)

に \* 助詞 格助詞 \* \* (ni particle Case\*\*)

F1:pred='Seoul',

F1:case='ni',

F2:obl=F1,

\* 2 -1D

行った いった 行く 動詞 \* 子音動詞 過去形 (itta 'went' iku Verb \* ConsonantStem pst)

F2:pred='iku',

F2:tns='pst',

F2:stmt='decl',

F2:style='plain'.

EOS

(a) The automatically annotated sentence

$$\left[ \begin{array}{l} \text{subj} \left[ \begin{array}{l} \text{pred 'Taro'} \\ \text{case ga} \end{array} \right] \\ \text{obl} \left[ \begin{array}{l} \text{pred 'Seoul'} \\ \text{case ni} \end{array} \right] \\ \text{pred 'iku' (subj, obl)} \\ \text{stmt 'decl'} \\ \text{style 'plain'} \\ \text{tense pst} \end{array} \right]$$

(b) The resulting f-structure

Figure 2: An example from the Kyoto Text Corpus: from syntactically annotated sentence to f-structure

(total 299 CPU seconds vs 200 CPU seconds to parse 560 sentences). The two approaches have the same goal: to provide a ranked list of LFG parses for a given input. The difference is in how this ranked list is derived, and how much manual effort is required. Furthermore, in Cahill, Maxwell, et al. (2008) it was shown that a simple pruning mechanism on the c-structure forests generated by the

*Aoife Cahill & Andy Way*

XLE parser could significantly reduce parsing time, while maintaining comparable accuracy.

## 8 Conclusion

This chapter has described methods based on LFG that permit accurate, robust, scalable, probabilistic LFG parsers and MT systems to be built from large collections of automatically annotated data. While this is commonplace nowadays, it was much less so 20–25 years ago.

LFG-DOP extends LFG by generalizing well-formed analyses to allow ill-formed and previously uncovered strings to be handled. LFG-DOT, a robust, hybrid model of translation based on LFG-DOP, was demonstrated to be able to solve ‘hard’ cases of translation that proved difficult for DOT and LFG-MT.

The range of work on automatic annotation of LFG grammars summarised here was an important step in ensuring scalability and robustness that is commonplace nowadays. Once large-scale treebanks could be generated via these techniques, competitive probabilistic parsers were built, and large-scale lexical resources were induced. However, most experiments carried out using LFG-DOP (and LFG-DOT) were relatively small-scale, but we sketch here a method for large-scale experimentation using the resources created via the techniques described in this paper.

As well as the important extension of the core LFG framework to account for probabilistic parsing, this seminal work also provided the foundations for the now commonplace task of large-scale deep linguistic LFG annotation. In sum, the work described in this chapter laid the foundations for multilingual annotation of treebanks, which in turn allowed competitive scalable parsing and MT models to be developed that are accepted as best practice today.

## Acknowledgements

We thank the anonymous reviewers for their comments, which have helped to improve this paper. We are grateful to Mary Dalrymple for her input regarding Pirahã. The second author is supported by the ADAPT SFI Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund.

## References

- Abeillé, Anne, François Toussanel & Martine Chéradame. 2004. *Corpus le monde: Annotations en constituants. Guide pour les correcteurs*. Tech. rep. Paris: Laboratoire de Linguistique Formelle, Université Paris Diderot.
- Arnold, Doug, Louisa Sadler, Ian Crookston & Andy Way. 1990. LFG and translation. In *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, 121–130. Austin, Texas.
- Bod, Rens. 1992. A computational model of language performance: Data-Oriented Parsing. In *COLING '92: proceedings of the 14th international Conference on Computational Linguistics*, vol. 3, 855–859. Nantes, France. DOI: 10.3115/992383.992386.
- Bod, Rens. 1998. *Beyond grammar: An experience-based theory of language*. Stanford, CA: CSLI Publications.
- Bod, Rens. 2000. An empirical evaluation of LFG-DOP. In *COLING 2000: the 18th Conference on Computational Linguistics*, vol. 1. Saarbrücken. DOI: 10.3115/990820.990830.
- Bod, Rens. 2007. Unsupervised syntax-based machine translation: The contribution of discontinuous phrases. In *Proceedings of the Machine Translation Summit XI*, 51–56. Copenhagen.
- Bod, Rens & Ronald M. Kaplan. 1998. A probabilistic corpus-driven model for lexical-functional analysis. In *ACL '98/COLING '98: Proceedings of the 36th annual meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, vol. 2, 145–151. Montréal. DOI: 10.3115/980451.980869.
- Bod, Rens & Ronald M. Kaplan. 2003. A Data-Oriented Parsing model for Lexical-Functional Grammar. In Rens Bod, Remko Scha & Ronald M. Kaplan (eds.), *Data-oriented parsing*. Stanford, CA: CSLI Publications.
- Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius & George Smith. 2002. The TIGER treebank. In *Proceedings of the 1st Workshop on Treebanks and Linguistic Theories*, 24–41.
- Burke, Michael. 2006. *Automatic annotation of the Penn-II Treebank with f-structure information*. Dublin: School of Computing, Dublin City University. (Doctoral dissertation).
- Burke, Michael, Aoife Cahill, Mairéad McCarthy, Ruth O'Donovan, Josef van Genabith & Andy Way. 2004. Evaluating automatic LFG f-structure annotation for the Penn-II treebank. *Research on Language and Computation* 2(4). 523–547. DOI: 10.1007/s11168-004-7428-y.

Aoife Cahill & Andy Way

- Burke, Michael, Olivia Lam, Aoife Cahill, Rowena Chan, Ruth O'Donovan, Adams Bodomo, Josef van Genabith & Andy Way. 2004. Treebank-based acquisition of a Chinese lexical-functional grammar. In *Proceedings of the 18th Pacific Asia conference on language, information and computation*, 161–172. Tokyo.
- Butt, Miriam. 1994. Machine translation and complex predicates. In *Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 94)*, 62–71. Vienna, Austria.
- Cahill, Aoife. 2004. *Parsing with automatically acquired, wide-coverage, robust, probabilistic LFG approximations*. Dublin: School of Computing, Dublin City University. (Doctoral dissertation).
- Cahill, Aoife, Michael Burke, Ruth O'Donovan, Stefan Riezler, Josef van Genabith & Andy Way. 2008. Wide-coverage deep statistical parsing using automatic dependency structure annotation. *Computational Linguistics* 34(1). 81–124. DOI: 10.1162/coli.2008.34.1.81.
- Cahill, Aoife, Michael Burke, Ruth O'Donovan, Josef van Genabith & Andy Way. 2004. Long-distance dependency resolution in automatically acquired wide-coverage PCFG-based LFG approximations. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL'04)*, 319–326. Barcelona. DOI: 10.3115/1218955.1218996.
- Cahill, Aoife, Martin Forst, Michael Burke, Mairéad McCarthy, Ruth O'Donovan, Christian Rohrer, Josef van Genabith & Andy Way. 2005. Treebank-based acquisition of multilingual unification grammar resources. *Journal of Research on Language and Computation; Special Issue on "Shared Representations in Multilingual Grammar Engineering"* 3(2–3). 247–279. DOI: 10.1007/s11168-005-1296-y.
- Cahill, Aoife, Martin Forst, Mairéad McCarthy, Ruth O'Donovan, Christian Rohrer, Josef van Genabith & Andy Way. 2003. Treebank-based multilingual unification-grammar development. In *Proceedings of the workshop on ideas and strategies for multilingual grammar development at the 15th European Summer School in Logic, Language and Information*, 17–24. Vienna.
- Cahill, Aoife, John T. Maxwell III, Paul Meurer, Christian Rohrer & Victoria Rosén. 2008. Speeding up LFG parsing using c-structure pruning. In *COLING 2008: Proceedings of the workshop on Grammar Engineering Across Frameworks*, 33–40. Manchester. DOI: 10.3115/1611546.1611551.
- Cahill, Aoife, Mairéad McCarthy, Josef van Genabith & Andy Way. 2002a. Automatic annotation of the Penn-Treebank with LFG F-structure information. In *LREC'02 workshop on linguistic knowledge acquisition and representation: Bootstrapping annotated language data*, 76–95. Las Palmas.
- Cahill, Aoife, Mairéad McCarthy, Josef van Genabith & Andy Way. 2002b. Parsing with PCFGs and automatic f-structure annotation. In Miriam Butt & Tracy



## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

- Holloway King (eds.), *Proceedings of the LFG '02 conference*, 76–95. Stanford, CA: CSLI Publications.
- Carroll, John, Edward Briscoe & Antonio Sanfilippo. 1998. Parser evaluation: A survey and new proposal. In *Proceedings of the international conference on language resources and evaluation*, 447–454. Granada.
- Carroll, John & Ted Briscoe. 2004. High precision extraction of grammatical relations. In Harry Bunt, John Carroll & Giorgio Satta (eds.), *New developments in parsing technology*, vol. 23 (Text, Speech and Language Technology). Dordrecht: Springer. DOI: 10.1007/1-4020-2295-6\_3.
- Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2(3). 113–124. DOI: 10.1109/tit.1956.1056813.
- Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris Publications. DOI: 10.1515/9783110884166.
- Chrupała, Grzegorz & Josef van Genabith. 2006. Improving treebank-based automatic LFG induction for Spanish. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '06 conference*. Stanford, CA: CSLI Publications.
- Civit, Montserrat. 2003. *Criterios de etiquación y desambiguación morfosintáctica de corpus en español*. Barcelona: Universitat Politècnica de Catalunya. (Doctoral dissertation).
- Collins, Michael. 1999. *Head-Driven statistical models for natural language parsing*. Philadelphia: University of Pennsylvania. (Doctoral dissertation).
- Cormons, Boris. 1999. *Analyse et désambiguisation: Une approche à base de corpus (Data-Oriented Parsing) pour les représentations lexicales fonctionnelles*. Université de Rennes. (Doctoral dissertation).
- Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III & Annie Zaenen (eds.). 1995. *Formal issues in Lexical-Functional Grammar*. Stanford, CA: CSLI Publications.
- Frank, Anette. 2000. Automatic f-structure annotation of treebank trees. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '00 conference*, 140–160. Stanford, CA: CSLI Publications.
- Futrell, Richard, Laura Stearns, Daniel L. Everett, Steven T. Piantadosi & Edward Gibson. 2016. A corpus investigation of syntactic embedding in Pirahã. *PLoS ONE* 11(3). DOI: 10.1371/journal.pone.0145289.
- Gazdar, Gerald & Chris Mellish. 1989. *Natural language processing in PROLOG: An introduction to computational linguistics*. Wokingham, UK: Addison-Wesley.
- Graham, Yvette & Josef van Genabith. 2012. Exploring the parameter space in statistical machine translation via f-structure transfer. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 254–270. Stanford, CA: CSLI Publications.

Aoife Cahill & Andy Way

- Guo, Yuqing. 2009. *Treebank-based acquisition of Chinese LFG resources for parsing and generation*. Dublin: School of Computing, Dublin City University. (Doctoral dissertation).
- Guo, Yuqing, Josef van Genabith & Haifeng Wang. 2007. Treebank-based acquisition of LFG resources for Chinese. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '07 conference*, 214–232. Stanford, CA: CSLI Publications.
- Hearne, Mary. 2005. *Data-oriented models of parsing and translation*. Dublin: School of Computing, Dublin City University. (Doctoral dissertation).
- Hearne, Mary & Khalil Sima'an. 2004. Structured parameter estimation for LFG-DOP. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova & Ruslan Mitkov (eds.), *Recent advances in natural language processing III: Selected papers from RANLP 2003*, vol. 260 (Current Issues in Linguistic Theory), 183–192. Amsterdam: John Benjamins. DOI: 10.1075/cilt.260.20hea.
- Hockenmaier, Julia & Mark Steedman. 2002. Generative models for statistical parsing with combinatory categorial grammar. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 335–342. Philadelphia.
- Kaplan, Ronald M. & Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 173–281. Cambridge, MA: The MIT Press. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 29–130).
- Kaplan, Ronald M., John T. Maxwell III, Tracy Holloway King & Richard Crouch. 2004. Integrating finite-state technology with deep LFG grammars. In *Proceedings of the workshop on combining shallow and deep processing for NLP at the European summer school on logic, language, and information (ESSLI)*.
- Kaplan, Ronald M., Klaus Netter, Jürgen Wedekind & Annie Zaenen. 1989. Translation by structural correspondences. In *Proceedings of the 4th conference of the European chapter of the ACL (EACL 1989)*, 272–281. DOI: 10.3115/976815.976852. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 311–330).
- Kaplan, Ronald M. & Jürgen Wedekind. 1993. Restriction and correspondence-based translation. In *Proceedings of the 6th conference of the European chapter of the ACL (EACL 1993)*, 193–202. DOI: 10.3115/976744.976768.
- Kay, Martin, Jean Mark Gawron & Peter Norvig. 1994. *Verbmobil: A translation system for face-to-face dialog*. Vol. 33. Stanford, CA: CSLI Publications.
- King, Tracy Holloway, Richard Crouch, Stefan Riezler, Mary Dalrymple & Ronald M. Kaplan. 2003. The PARC 700 Dependency Bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora, held at the 10th*

## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

*Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*. Budapest.

- Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit X*, 79–86. Phuket, Thailand.
- Koehn, Philipp, Franz Josef Och & Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 127–133. Edmonton, Canada. DOI: [10.21236/ada461156](https://doi.org/10.21236/ada461156).
- Kurohashi, Sadao & Makoto Nagao. 1997. Kyoto Daigaku text corpus project. In *Proceedings of the third annual meeting of the association of applied natural language processing*, 115–118. In Japanese.
- Lappin, Shalom, Igal Golan & Mori Rimón. 1989. *Computing grammatical functions from configurational parse trees*. Technical Report 88.268. Haifa: IBM Israel.
- Maamouri, Mohamed & Ann Bies. 2004. Developing an Arabic treebank: Methods, guidelines, procedures, and tools. In *Proceedings of the workshop on computational approaches to Arabic script-based languages*, 2–9. Geneva. DOI: [10.3115/1621804.1621808](https://doi.org/10.3115/1621804.1621808).
- Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz & Britta Schasberger. 1994. The Penn treebank: Annotating predicate argument structure. In *Proceedings of the workshop on human language technology*, 114–119. Plainsboro, NJ. DOI: [10.3115/1075812.1075835](https://doi.org/10.3115/1075812.1075835).
- McCarthy, Mairéad. 2003. *Design and evaluation of the linguistic basis of an automatic f-structure annotation algorithm for the Penn-II Treebank*. Dublin: School of Computing, Dublin City University. (MA thesis).
- Miyao, Yusuke, Takashi Ninomiya & Jun'ichi Tsujii. 2003. Probabilistic modeling of argument structures including non-local dependencies. In *Proceedings of the conference on recent advances in natural language processing (RANLP 2003)*, 285–291. Borovets, Bulgaria.
- O'Donovan, Ruth. 2006. *Automatic extraction of large-scale multilingual lexical resources*. Dublin: School of Computing, Dublin City University. (Doctoral dissertation).
- O'Donovan, Ruth, Michael Burke, Aoife Cahill, Josef van Genabith & Andy Way. 2004. Large-scale induction and evaluation of lexical resources from the Penn-II Treebank. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics*, 368–375. Barcelona. DOI: [10.3115/1218955.1219002](https://doi.org/10.3115/1218955.1219002).

Aoife Cahill & Andy Way

- O'Donovan, Ruth, Aoife Cahill, Josef van Genabith & Andy Way. 2005. Automatic acquisition of Spanish LFG resources from the CAST3LB treebank. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '05 conference*, 334–352. Stanford, CA: CSLI Publications.
- Oya, Masanori & Josef van Genabith. 2007. Automatic acquisition of Lexical-Functional Grammar resources from a Japanese dependency corpus. In *Proceedings of the 21st Pacific Asia conference on language, information and computation*, 375–384. Seoul.
- Poutsma, Arjen. 2000. Data-oriented translation. In *COLING 2000: the 18th Conference on Computational Linguistics*, vol. 2, 635–641. Saarbrücken. DOI: 10.3115/992730.992738.
- Rehbein, Ines & Josef van Genabith. 2009. Automatic acquisition of LFG resources for German – As good as it gets. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 480–500. Stanford, CA: CSLI Publications.
- 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 (ACL '02)*. Philadelphia.
- Riezler, Stefan & John T. Maxwell III. 2006. Grammatical machine translation. In *Proceedings of the human language technology conference of the NAACL, main conference*, 248–255. New York City: Association for Computational Linguistics. DOI: 10.3115/1220835.1220867.
- Rimell, Laura, Stephen Clark & Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 conference on Empirical Methods in Natural Language Processing*, 813–821. Singapore. DOI: 10.3115/1699571.1699619.
- Sadler, Louisa & Henry Thompson. 1991. Structural non-correspondence in translation. In *Proceedings of the 5th conference of the European chapter of the ACL (EACL 1991)*, 293–298. Berlin. DOI: 10.3115/977180.977231.
- Sadler, Louisa, Josef van Genabith & Andy Way. 2000. Automatic f-structure annotation from the AP treebank. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '00 conference*, 226–243. Stanford, CA: CSLI Publications.
- Schluter, Natalie. 2011. *Treebank-based deep grammar acquisition for French probabilistic parsing resources*. Dublin: School of Computing, Dublin City University. (Doctoral dissertation).

## 5 Treebank-driven Parsing, Translation and Grammar Induction using LFG

- Schluter, Natalie & Josef van Genabith. 2007. Preparing, restructuring, and augmenting a French treebank: Lexicalised parsers or coherent treebanks? In *Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation*, 200–209. Melbourne.
- Schluter, Natalie & Josef van Genabith. 2008. Treebank-based acquisition of LFG parsing resources for French. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, 2909–2916. Marrakech.
- Sima'an, Khalil. 1997. An optimized algorithm for Data-Oriented Parsing. In Ruslan Mitkov & Nicolas Nicolov (eds.), *Recent advances in natural language processing: Selected papers from RANLP '95* (Current Issues in Linguistic Theory), 35–46. Amsterdam: John Benjamins. DOI: [10.1075/cilt.136.05sim](https://doi.org/10.1075/cilt.136.05sim).
- Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister & Kathrin Beck. 2006. *Stylebook for the Tübingen treebank of written German (TüBa-D/Z)*. Tübingen: Universität Tübingen.
- Tounsi, Lamia, Mohammed Attia & Josef van Genabith. 2009a. Automatic treebank-based acquisition of Arabic LFG dependency structures. In *Proceedings of the EACL 2009 workshop on computational approaches to Semitic languages*, 45–52. Athens, Greece. DOI: [10.3115/1621774.1621783](https://doi.org/10.3115/1621774.1621783).
- Tounsi, Lamia, Mohammed Attia & Josef van Genabith. 2009b. Parsing Arabic using treebank-based LFG resources. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 583–586. Stanford, CA: CSLI Publications.
- van Genabith, Josef, Anette Frank & Michael Dorna. 1998. Transfer constructors. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '98 conference*, 190–205. Stanford, CA: CSLI Publications.
- van Genabith, Josef, Louisa Sadler & Andy Way. 1999. Deriving an LFG from a treebank resource. In *Proceedings of the ATALA international workshop on treebanks*, 107–114. Paris.
- van Genabith, Josef, Andy Way & Louisa Sadler. 1999a. Data-driven compilation of LFG semantic forms. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora (LINC-99)*, 69–76. Bergen.
- van Genabith, Josef, Andy Way & Louisa Sadler. 1999b. Semi-automatic generation of f-structures from treebanks. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '99 conference*, 19–21. Stanford, CA: CSLI Publications.
- Way, Andy. 1999. A hybrid architecture for robust MT using LFG-DOP. *Journal of Experimental and Theoretical Artificial Intelligence, Special Issue on Memory-Based Learning* 11. 441–471. DOI: [10.1080/095281399146490](https://doi.org/10.1080/095281399146490).

*Aoife Cahill & Andy Way*

- Way, Andy. 2001. *A hybrid architecture for robust MT*. Colchester, UK: University of Essex. (Doctoral dissertation).
- Way, Andy. 2003. Machine translation using LFG-DOP. In Rens Bod, Remko Scha & Khalil Sima'an (eds.), *Data-oriented parsing*, 359–384. Stanford, CA: CSLI Publications.
- Way, Andy, Ian Crookston & Jane Shelton. 1997. A typology of translation problems for Eurotra translation machines. *Machine Translation* 12. 323–374.
- Xia, Fei. 1999. Extracting tree adjoining grammars from bracketed corpora. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS-99)*, 398–403. Peking.
- Xue, Nianwen, Fu-Dong Chiou & Martha Palmer. 2002. Building a large-scale annotated Chinese corpus. In *COLING 2002: the 19th International Conference on Computational Linguistics*. Taipei. DOI: 10.3115/1072228.1072373.
- Zaenen, Annie & Ronald M. Kaplan. 1995. Formal devices for linguistic generalizations: West Germanic word order in LFG. In Jennifer S. Cole, Georgia M. Green & Jerry L. Morgan (eds.), *Linguistics and computation*, 3–27. Stanford, CA: CSLI Publications. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 215–240).

## **Part IV**

# **Language families and regions**





# Chapter 6

## LFG and Finno-Ugric languages

Tibor Laczkó

Károli Gáspár University of the Reformed Church in Hungary

The chapter discusses some salient, sometimes competing, LFG analyses of a variety of (morpho-)syntactic phenomena in Finno-Ugric languages, with occasional glimpses at alternative generative approaches and at some related phenomena in languages belonging to Samoyedic, the other major branch of Uralic languages. It concentrates on clausal c-structure representational issues, verbal modifiers, focused constituents, negation, copula constructions, argument realization, subject-verb agreement, differential object marking, evidentiality and a set of noun phrase phenomena related to event nominalization. It argues that LFG provides an appropriate and suitably flexible formal apparatus for a principled analysis of all the phenomena in all the Finno-Ugric languages discussed here. In addition, it shows that the analysis of some of these phenomena can also contribute to LFG-internal theorizing.

### 1 Introduction

#### 1.1 General remarks on Finno-Ugric languages

Finno-Ugric is one of the two branches of Uralic, the other branch being Samoyedic. In Figure 1 we show the major branches of the Uralic family tree and those leaves (languages) that are discussed, or at least mentioned, in this chapter. This figure capitalizes on the general remarks in the introductory chapter of Miestamo et al. (2015) on the representation of the Finno-Ugric branch.<sup>1</sup> We use the names of the individual languages as they appear in that volume.<sup>2</sup> The authors point out that,

<sup>1</sup>We are thankful to Anne Tamm for helpful discussions of certain family tree issues.

<sup>2</sup>Several languages in this figure are also referred to by alternative names in some other works, e.g. Khanty = Ostyak, Mansi = Vogul, Udmurt = Votyak, Mari = Cheremis; see the discussion



*Tibor Laczkó*

although there are several alternative approaches to this branch, most of them share the view that the following language groups are valid genealogical units: Samoyedic, Ugric, Permic, Mari, Mordvin, Saamic and Finnic. However, the details of the relationships among certain languages are subject to variation across these competing approaches.<sup>3</sup>

For the sake of a complete picture, we have included the Samoyedic branch as well. In the Northern branch there are two major sub-branches: Enets-Nenets and Nganasan. From the Enets-Nenets sub-branch Tundra Nenets will be discussed and compared with some Finno-Ugric languages in Section 7.1.2 with respect to differential object marking. The only living representative of the Southern branch is Selkup, also mentioned in Section 7.1.2. Saamic languages also have a variety of sub-branches. From these languages Inari Saami will be discussed in Section 5.2 on copula constructions and in Section 7.1.1 on subject-verb agreement.

As regards the geographical distribution of the languages indicated in Figure 1, Estonian is primarily spoken in Estonia, Hungarian is spoken in Hungary, Finnish and Inari Saami are spoken in Finland, and all the other languages are spoken in Russia.

Several languages belonging to the Finno-Ugric branch of Uralic languages have a considerable number of properties that have contributed to linguistic research in LFG. On the one hand, these phenomena provide empirical or typological evidence for theoretical generalizations. On the other hand, they exhibit cases in which LFG is well-suited for the development of principled analyses. Such phenomena include, but are not limited to, discourse-functionality, negation, *wh*-questions, copular clauses, particle-verb constructions, event nominalization, possessive constructions, the nature and inventory of grammatical functions, evidentiality, rich inflectional morphology, partitives, duals and complex agreement patterns.

In this chapter we can only concentrate on those phenomena in Finno-Ugric languages that have been analyzed in an LFG framework in such a way that the summary of the given analysis within the limitations of space serves the purposes of the chapter, as outlined in the foregoing paragraph. Consequently, this determines which languages appear in the chapter. Given that Hungarian is the most intensively and extensively researched Finno-Ugric language in LFG,<sup>4</sup> the dis-

---

of Dalrymple & Nikolaeva (2011) in Section 7.1.2, for instance. When we cite authors, we keep the version of the name of a language that they use.

<sup>3</sup>For a recent, fundamentally similar Uralic family tree representation indicating all the languages (including those that are extinct by now), see Maticsák (2020).

<sup>4</sup>For introductions to LFG in Hungarian, see Laczkó (1989) and Komlósy (2001). The following

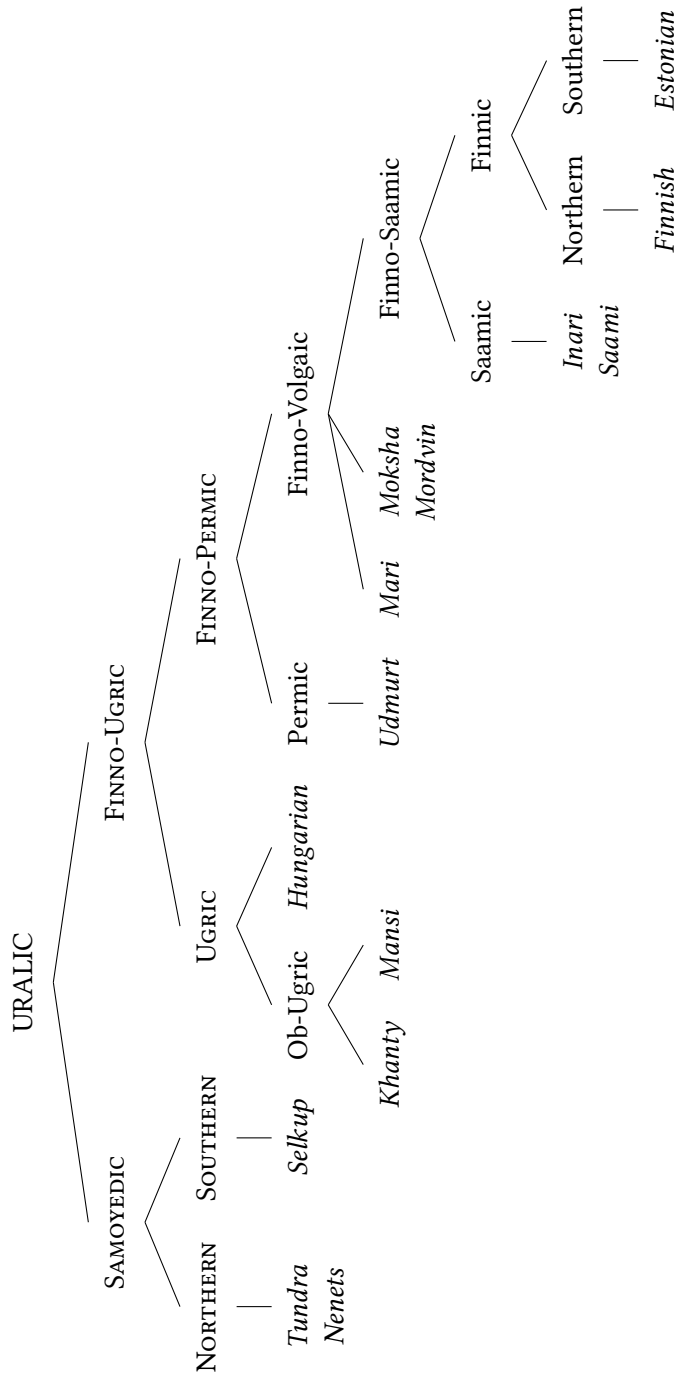


Figure 1: The (simplified) family tree of Uralic languages

*Tibor Laczkó*

cussions of LFG analyses of Hungarian phenomena outnumber the discussions of phenomena in other Finno-Ugric languages. For further information on related and additional phenomena and other Uralic or Finno-Ugric languages, the interested reader is referred to the following comprehensive sources: Abondolo (1998), Dryer & Haspelmath (2013), Miestamo et al. (2015) and Groot (2017).<sup>5</sup> The online journal *Finno-Ugric Languages and Linguistics* (<http://full.btk.ppke.hu>) regularly contains generative papers on Finno-Ugric languages.<sup>6</sup> In addition, Tamm & Vainikka (2018) present an overview of generative works on Finnish and Estonian syntax.<sup>7</sup>

As regards comprehensive analyses of several phenomena in Hungarian, Laczkó (to appear) offers a synthesis of his earlier LFG(-XLE)<sup>8</sup> accounts of the following phenomena in Hungarian finite clauses: sentence structure, verbal modifiers, operators, negation and copula constructions. He posits all this in the context of a critical overview of alternative Chomskyan and lexicalist approaches to these phenomena. Tamm (2004c) develops a comprehensive LFG approach to the relations between Estonian aspect, verbs and case.

The following databases on Uralic languages are useful resources about their syntactic properties: the Selkup and Kamas corpora at <https://www.slm.uni-hamburg.de/inel/>, the typological database of Ugric languages at <http://en.utdb.nullpoint.info/> and the Uralic language typological data set at <https://bedlan.net/data/uralic-language-typological-data-set/>.

---

works also have introductory sections to LFG: Szabó (2017) in Hungarian and Tamm (2004a) in Estonian.

<sup>5</sup>In Section 9 we make brief references to additional works on Uralic languages in general and Finno-Ugric languages in particular that we cannot discuss here for limitations of space.

<sup>6</sup>See, for instance, Brattico (2019) on Finnish word order, É. Kiss (2020) on pronominal objects in Ob-Ugric, and Asztalos (2020) on focus in Udmurt.

<sup>7</sup>In her review, Anne Tamm has kindly provided the following information about the history of syntactic research on Estonian. ‘For a long while since the mid-20<sup>th</sup> century, there was more work on Estonian syntax than on Finnish syntax. Keeping abreast with western mainstream linguistics in the 60s, 70s and early 80s resulted in numerous formal syntactic works and a tradition of understanding syntax that is, in spirit, rather similar to LFG approaches. Rätsep (1978), for instance, is a lexicalist analysis of patterns of argument structures and their alternations; this work has certainly been influential in the context Uralic syntax. Tamm (2012c) provides an overview of the treatment of verb classes in this and related works, these early generative-style lexicalist works are available in Estonian only [...]. Almost all LFG work on Estonian expands that work in some way.’

<sup>8</sup>In his XLE work he further develops Laczkó & Rákosi’s (2008–2019) implemented Hungarian grammar.

## 1.2 The structure of the chapter

In accordance with the scopes of LFG works on Finno-Ugric languages, the significantly larger part of this chapter (Section 2–Section 7) concentrates on the investigation of clausal phenomena, and this is followed by the discussion of salient LFG analyses of some noun phrase phenomena (Section 8). In Section 2 we discuss clausal c-structure representational issues by focusing on a variety of LFG approaches to Hungarian. In Section 3 we concentrate on verbal modifiers in Hungarian and Estonian in general and on their radically different relations to focus in these languages in particular. In Section 4 we offer a brief overview of an LFG-XLE analysis of negation in Hungarian by also pointing out its potential contribution to the treatment of negation phenomena cross-linguistically. In Section 5 we discuss LFG accounts of copula constructions in Hungarian, Inari Saami and Finnish. In Section 6 we deal with LFG treatments of some aspects of argument realization in Finnish and Estonian. In Section 7 we concentrate on a selection of morphosemantic phenomena: (i) subject-verb agreement in Inari Saami and Finnish; (ii) differential object marking in Uralic with particular attention to Finno-Ugric languages; (iii) the grammaticalized expression of evidentiality in Udmurt and Estonian. In Section 8 we present a summary of a variety of LFG approaches to noun phrase phenomena in Hungarian: (i) c-structure issues; (ii) event nominalization, and we add a short section on the morpho-syntax of possessive noun phrases in Finnish and Hungarian. In Section 9 we make brief references to additional relevant LFG(-related) works on Finno-Ugric languages that space limitations have prevented us from discussing. In Section 10 we conclude.

## 2 C-structure representation in clauses

Groot (2017) presents a very useful tabular comparison of the major word order properties of 21 Uralic languages. In Table 1 we present the parts of his table that are relevant for our current purposes.

As the table shows, in these languages word order is predominantly free (except for Enets and Nenets). The two major patterns are SVO and SOV with roughly the same frequency. In seven languages there is a designated preverbal focus position (and in one of them, Komi, there is an additional postverbal Foc position). In three languages the Foc position is clause final. This is the general word order picture. Below we fundamentally concentrate on Hungarian because several alternative LFG c-structure analyses have been proposed for this language. In addition, we make some comparative remarks on Finnish and Estonian.

*Tibor Laczkó*

Table 1: Word order properties of 21 Uralic languages (part of Table 11, Groot 2017: 548)

Language	word order	major pattern	focus position
Finnish	free	SVO	
Estonian	free	SVO	clause final
Votic	free	SVO	
Ingrian	free	SVO	clause final
Veps	free	SVO	clause final
Karelian	free	SVO	
South Saami	free	SOV/SVO	
North Saami	free	SVO	
Skolt Saami	free	SVO	
Erzya	free	SVO	
Mari	free	SOV	Foc V
Komi	free	SOV/SVO	Foc V / V Foc
Udmurt	free	SOV	Foc V
Hungarian	free	SOV/SVO	Foc V
Khanty	free	SOV	Foc V
Mansi	free	SOV	
Nenets	not free	strict SOV	Foc V
Enets	not free	strict SOV	Foc V
Nganasan	free	SOV	
Selkup	free	SOV	
Kamas	free		

Hungarian is a classic example of a discourse configurational language: see É. Kiss (1995), for instance.<sup>9</sup> The crucial empirical generalizations about Hungarian sentence structure are as follows. The fundamental sentence articulation

<sup>9</sup>On sentence structure and discourse-functionality in Finnish in non-LFG frameworks, see Vilkuna (1995) and Brattico (2019), for instance. According to Vilkuna (1995), there is a preverbal K (contrast) and also a T (topic) position in Finnish. While fundamentally these two positions are also available in Estonian, on the basis of their experimental and corpus investigation, Sahkai & Tamm (2018b, 2019) claim that other types of constituents can also occur in the preverbal domain. While Hungarian exhibits strong discourse-configurationality, Estonian is only weakly discourse-configurational: see Sahkai & Tamm (2018b: 416–417). Hiietam (2003) argues that topic is to be defined semantically and not configurationally in this language. In addition, Estonian is the only Uralic language with V2, and its V2 is prosodic: see Sahkai & Tamm (2018a). Tael (1988) claims that the focus position is at the end of the clause in Estonian.

## 6 LFG and Finno-Ugric languages

is topic-predicate (also called topic-comment in a variety of approaches). In the topic field, the order of topics and sentence adverbs is free. In the preverbal domain, quantifiers follow the topic field. In neutral sentences<sup>10</sup> there is a designated immediately preverbal position for a special constituent type: ‘verbal modifier’ (VM). This is a conventionally used cover term for a range of radically different categories sharing the syntactic property of occupying this designated preverbal position. Preverbs (also known as verbal particles or coverbs),<sup>11</sup> bare nouns, designated XP arguments, etc. are all assumed to be VMs. Basically, the word order of postverbal elements is also free. In a non-neutral sentence the (heavily stressed) focused constituent occupies the immediately preverbal position, and, as a consequence, the VM has to occur postverbally, i.e. the VM and the focus are in complementary distribution. How to capture this complementarity is a crucial cross-theoretical issue. The two salient solutions are as follows. (i) There is only a single designated preverbal position for which focused constituents and VMs compete. (ii) There are two distinct positions for the two elements: focus and VM. In this approach it needs to be explained why these two elements cannot co-occur.

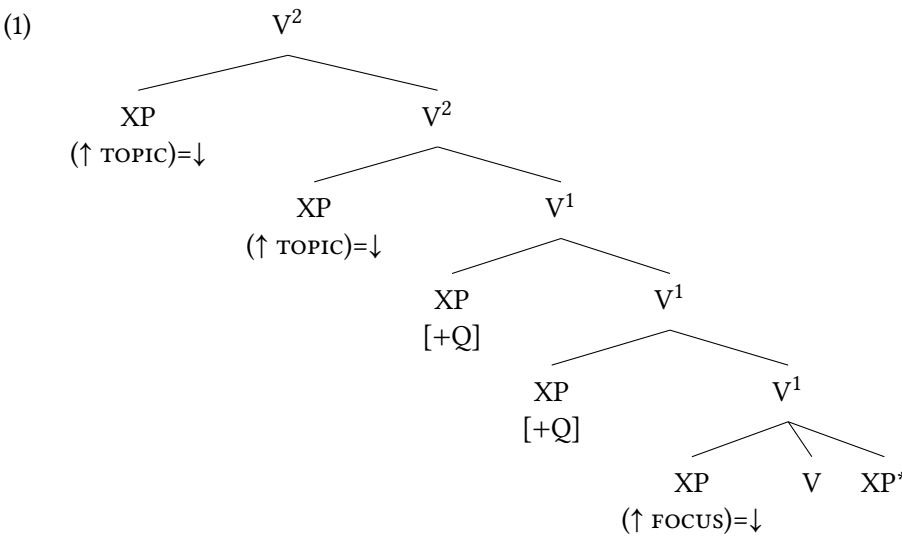
Börjars et al. (1999) offer some general considerations against functional projections like TopP and FocP (*à la* GB and MP) for languages like Hungarian and some hints at a possible LFG alternative with an extended verbal projection in which word order regularities are capturable by dint of Optimality Theoretic (OT) style constraints. They claim that the assumption that discourse functions are not necessarily associated with the specifier positions of functional projections allows an analysis of Hungarian in which quantifier phrases and topics are positioned within an extended verbal projection, avoiding the postulation of functional projections without heads. They propose that Hungarian sentences are VP projections, as in (1),<sup>12</sup> and they suggest that the immediately preverbal occurrence of the focused constituent should be captured in terms of OT constraints. In this work, there is no discussion of VMs and their complementarity with focused phrases.

<sup>10</sup>The standard description of a neutral sentence is that it does not contain negation or focus, it is not a *wh*-question, and it has level prosody.

<sup>11</sup>Other Ob-Ugric languages have developed verbal particles to a lesser extent, see Zsirai (1933). For more information on Uralic (aspectual) verbal prefixation and verbal particles, see Kiefer & Honti (2003). For an analysis of Estonian sentence-final particles with focus, see Tamm (2004c: 224–242), discussed in Section 3.2.

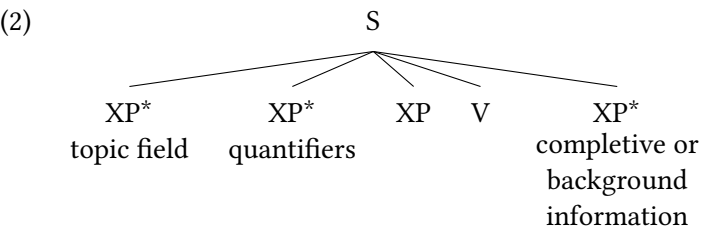
<sup>12</sup>The superscripts in V<sup>1</sup> and V<sup>2</sup> indicate bar-levels.

*Tibor Laczkó*



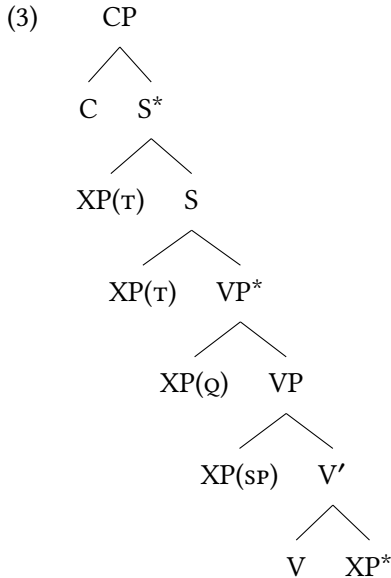
Adopting the basic representational assumptions and ideas of Börjars et al. (1999), in their OT-LFG framework, Payne & Chisarik (2000) develop an analysis of Hungarian preverbal syntactic phenomena: the complementarity of constituent question expressions, focused constituents, the negative marker and verbal modifiers.

Gazdik (2012), capitalizing on Gazdik & Komlósy (2011), outlines an LFG analysis of Hungarian finite sentence structure, predominantly driven by discourse functional assumptions and considerations. She postulates two sentence structure types, and she assumes that both structures are available to both neutral (N) and non-neutral (NN) sentences, which are distinguished by their different prosodic behaviours. (2) shows one of the two structures. Here the immediately preverbal XP has a presentational-focus-like function in N sentences and the standard identificational focus function in NN sentences. The other structure differs in one important respect: the preverbal element is a VM, and the VM and the verb are dominated by V'. The VM receives the usual phonological-word-initial stress in N sentences and the focus stress in NN sentences.





Laczkó (2014b), after a detailed critical overview of previous LFG approaches, postulates the skeletal sentence structure in (3).<sup>13</sup> He argues against assuming an IP for the structural-categorical representation of Hungarian sentences and he argues for S as the core category.<sup>14</sup> He proposes a CP/S alternative that is closest in spirit to É. Kiss' (1992) special GB approach.<sup>15</sup>



Adopting one of the most crucial aspects of É. Kiss's (1992) analysis, he assumes that vms and focused constituents target the Spec,VP position. He employs disjunctive functional annotations to capture this preverbal complementarity.<sup>16</sup>

Consider the following generalization. 'The daughters of S may be subject and predicate' (Bresnan 2001: 112). In his analysis, Laczkó proposes that this generalization should be modified in the following way.

- (4) The daughters of S may be subject/topic and predicate.

He points out that this modification receives independent support from the following rule from Bresnan & Mchombo (1987).<sup>17</sup>

<sup>13</sup>In (3) τ stands for topic (position), Q stands for quantifier (position), sp stands for the specifier position. S\* and VP\* encode the possibly iterative left-adjunction of XP(τ) and XP(Q) to S and VP, respectively.

<sup>14</sup>In LFG IP and S are taken to be parametric options in Universal Grammar.

<sup>15</sup>For a comparison of these GB and LFG approaches, see Laczkó (2020).

<sup>16</sup>For details and the discussion of what other elements are assumed to compete for the Spec,VP position, see Section 3.1 and Section 4.

<sup>17</sup>On the basis of (5), *subject and/or topic* is even more appropriate than *subject/topic* in (4).

Tibor Laczkó

$$(5) \quad S \quad \longrightarrow \quad \left( \begin{array}{c} \text{NP} \\ (\uparrow \text{SUBJ})=\downarrow \end{array} \right), \left( \begin{array}{c} \text{NP} \\ (\uparrow \text{TOPIC})=\downarrow \end{array} \right), \left( \begin{array}{c} \text{VP} \\ (\uparrow=\downarrow) \end{array} \right)$$

Laczkó argues that a VP can contain a subject if the XP in [<sub>S</sub> XP VP] is a topic. This requires all other occurrences of VP to be subjectless. In this scenario, the following three parametric options seem to emerge across languages: (i) strictly VP-external subject, as in English; (ii) VP-internal subject in a designated position, as in Russian<sup>18</sup>; (iii) VP-internal subject without a designated position, see Hungarian.

This section has demonstrated that LFG provides a suitably flexible formal apparatus by the help of which the sentence structures of typologically different languages can be described in a principled manner with respect to discourse functional configurationality.

### 3 Verbal modifiers and focus

In this section we discuss analyses of verbal modifiers in Hungarian (Section 3.1) and Estonian (Section 3.2).

#### 3.1 Hungarian

As has been pointed out in Section 2, the crucial (cross-)theoretical question to address in the case of Hungarian is how to account for the preverbal complementarity of focused constituents and verbal modifiers. Compare the examples in (6). (6a) is a neutral sentence and the *vm oda* ‘to.there’, which is categorially a preverb, immediately precedes the verb. By contrast, (6b) is a non-neutral sentence, and in it the *vm* can neither precede nor follow the focused constituent (in SMALLCAPS) in the preverbal domain.

(6) Hungarian:

- a. János      minden-t      oda adott Mari-nak.  
    John.NOM everything-ACC *vm* gave Mary-DAT  
    ‘John gave everything to Mary.’
- b. János      minden-t      (\*oda) MARI-NAK (\*oda) adott oda.  
    John.NOM everything-ACC *vm*    Mary-DAT *vm*    gave *vm*  
    ‘John gave everything TO MARY.’

---

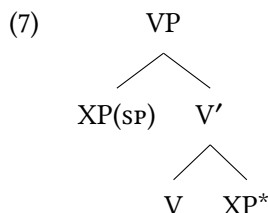
<sup>18</sup>See King (1995), for instance.

## 6 LFG and Finno-Ugric languages

The cross-theoretic question is whether we should assume that the two constituents fight for one and the same syntactic position or that they occupy two distinct positions. With the salient exception of É. Kiss (1992, 1994), the GB/MP mainstream assumes two distinct positions, and employs a variety of principles that block the simultaneous occurrence of constituents in these positions: see Brody (1990) and É. Kiss (2004), for instance, and also see Laczkó (to appear) for a comparative overview of different analyses of the complementarity of focused constituents and verbal modifiers in Hungarian.

Several LFG approaches have a similar view, see Ackerman (1987, 1990), Payne & Chisarik (2000), Mycock (2006, 2010), the basic idea being that vms get semantically and morphologically incorporated into the verb.<sup>19</sup> In Section 2 we also pointed out that Gazdik (2012) has a special proposal. She employs two distinct sentence structures, both having neutral and non-neutral versions. The main point here is that the basic VM vs. focus contrast is treated in two different structural dimensions. Thus, this can be regarded as an extreme instance of assuming that the two elements do not fight for the same syntactic position.

By contrast, Laczkó (2014b) argues that focus constituents (ordinary foci, the immediately preverbal *wh*-phrases and negated constituents)<sup>20</sup> and vms (of various types) target the same Spec,VP position, hence their complementarity. In (7) we repeat the relevant part of his overall sentence structure shown in (3) in Section 2.



Laczkó (2014b) employs disjunctive functional annotations to capture the complementarity of the elements he assumes to compete for this position.

As we pointed out in Section 2, vms come in several varieties: preverbs, idiom chunks, secondary predicates, designated reduced or full arguments. Preverbs are the central and theoretically by far the most challenging members of this

<sup>19</sup> At first sight, it can be taken to be a supporting fact that the vms of the preverb type and the verb make up one phonological word, i.e. it is only (the first syllable of) the preverb that receives word-initial stress. However, even XP vms follow the same pattern (in which the following verb loses its word-initial stress).

<sup>20</sup> On the details of negation in Hungarian, see Section 4.

*Tibor Laczkó*

heterogeneous group, because their combination with the finite verb, often called particle-verb construction (pvc), exhibits both lexical and syntactic properties (and the former motivate the incorporation analysis). Their most salient lexical characteristics are as follows. The preverb can affect the argument structure of the main verb, pvc's are often non-compositional, and both non-compositional and compositional pvc's can undergo productive derivational processes like event nominalization. However, the preverb and the main verb are strictly separable syntactically under clearly definable circumstances. For instance, as exemplified in (6) above, a focused constituent, as a rule, immediately precedes the main verb, and in such cases the preverb must occur postverbally.

In several recent LFG approaches, for instance Forst et al. (2010), Laczkó & Rákosi (2011), Rákosi & Laczkó (2011), Laczkó (2013a) and Laczkó (2014b), it is assumed that preverbs and other types of vms uniformly occupy a distinct preverbal syntactic position (typically Spec,VP), as opposed to the vm-incorporation analysis, which is primarily motivated by the preverbal complementarity of vms, focused and *wh*-constituents.

Forst et al. (2010) propose an LFG-XLE treatment of a variety of particle-verb constructions in English, German and Hungarian. Their main claim is that non-compositional and non-productive pvc's should be treated radically differently from compositional and productive pvc's. The former are best analyzed along lexical lines with the help of XLE's *CONCATENATION* device. By contrast, the authors argue that the productive pvc types call for a syntactic treatment. One of the most important motivations for this sharp distinction is that productive pvc's can be analyzed 'on the fly', i.e. automatically and straightforwardly, in the syntax, without previously and lexically encoding them. Their solution is complex predicate formation in the syntax by applying XLE's *RESTRICTION* operator.<sup>21</sup>

Laczkó & Rákosi (2011) and Rákosi & Laczkó (2011) explore the tenability and implementational applicability of the approach proposed by Forst et al. (2010) by each developing an LFG-XLE analysis of two different pvc types. Laczkó & Rákosi (2013) posit this approach in a cross-linguistic and cross-theoretical context. As opposed to previous LFG accounts, Laczkó (2013a) argues that compositional pvc's should also be treated lexically in a manner similar to the treatment of non-compositional pvc's. He points out that one of the advantages of this uniform lexical treatment is that classical LFG's view of the distribution of labour between the lexical and the syntactic components of grammar can be maintained, at least in this domain. He also shows how various morphological processes (often consecutively) involving pvc's can be handled (e.g. causativization, nominalization,

---

<sup>21</sup>For formal details, see Forst et al. (2010).

and preverb reduplication), which may cause potential problems for a syntactic analysis of compositional pvc's.

Laczko (2014b) captures the preverbal complementarity of focused constituents and vms by assuming that they fight for the same Spec,VP position. He encodes this by associating the disjunctive sets of annotations in (8) with this position. The first disjunct of the main disjunction says that a constituent bearing any grammatical function can have the focus discourse function. The second disjunct handles vms. Laczko employs XLE's CHECK feature device here.<sup>22</sup>

$$(8) \quad \left\{ \begin{array}{l} (\uparrow \text{ GF}) = \downarrow \\ (\uparrow \text{ FOCUS}) = \downarrow \\ | (\downarrow \text{ CHECK\_VM}) =_c + \\ \{ \uparrow = \downarrow \\ | (\uparrow \text{ GF}) = \downarrow \} \end{array} \right\}$$

The CHECK feature in (8) is used for all types of vms. It requires the presence, in Spec,VP, of an element lexically marked with the defining counterpart of this feature. Preverbs are intrinsically associated with this feature, i.e. in their lexical forms they are associated with the defining member of the CHECK\_VM feature pair, and they receive the functional (co-)head annotation, see the first disjunct in the second major disjunct. All the other types of vms are specified for this status by individual verbs. It depends on the verb whether it selects a vm, and, if so, which argument (bearing any subcategorized grammatical function) will be singled out, see the second disjunct in the second major disjunct.<sup>23</sup>

Laczko (2014a) outlines an LFG analysis of a variety of vms other than preverbs: bare nouns,<sup>24</sup> OBL XP arguments, xCOMP arguments and idiom chunks.

<sup>22</sup>The essence of this device is that CHECK features come in pairs: there is a defining equation and it has a constraining equation counterpart. These CHECK feature pairs, which can be used both in c-structure representations and lexical forms, can ensure that two elements will occur together in a particular configuration or a particular element occurs in a particular position. The CHECK feature in (8) is of the latter type.

<sup>23</sup>Laczko also assumes that a *wh*-phrase (or, in multiple *wh*-questions the immediately preverbal *wh*-phrase) also fights for the Spec,VP position, so he adds another disjunction to (8) to capture this, by using additional (interrogative) CHECK features: for details, see Laczko (2014b). In addition, he assumes that negated constituents also occupy this position. Furthermore, he postulates that in the type of predicate negation in which there is no focused constituent, the negative marker also targets this position. Therefore, he adds two more disjuncts, see Section 4.

<sup>24</sup>Viszket (2004) offers a detailed empirical description of a whole range of bare noun phrases in Hungarian. In neutral sentences these constituents can only occur immediately preverbally, in the vm position. In her LFG account of the syntax of bare noun phrases, Vizsket adopts Laczko's (1995; 2000b) [+vm] feature and she also introduces a special [ $\bullet$ vm] feature. Her new

*Tibor Laczkó*

The crucial aspect of this analysis is that in the lexical form of the verb taking any one of these VM types it is specified that either the verb occurs in a sentence containing a focused constituent or else its designated complement must occupy the Spec,VP position.<sup>25</sup>

### 3.2 Estonian

Tamm (2004c) presents a detailed description of pvc's in Estonian, and she outlines an LFG analysis. She points out that Estonian separable particles are basically comparable to their Hungarian counterparts, the most important difference being that aspectual particles typically occupy the clause final focus position. Tamm distinguishes three basic uses of Estonian particles, and she discusses the particle *ära*, which can perform all the three functions. Consider her examples.

- (9) Directional (deictic) use of *ära*, Estonian:

*ära*    *veerema*  
away roll  
'roll away'

Tamm points out that verbs combining with *ära* in this use have an implicit path argument that is only optionally realized overtly. The closest Hungarian counterpart is *el* 'away' (as in *el-gurul* 'roll away').

- (10) Completive use of *ära*, Estonian:

*Naaber*    *suri*            *ära*.  
neighbour die.PST.3SG ÄRA  
'The neighbour died.'

---

feature, when associated with a predicate in its lexical form, bans the occurrence of a bare NP in the VM position; practically, it prevents such a constituent from occurring in neutral sentences. Viszket identifies seven major types of predicates that need to be provided with this feature in their lexical forms. For instance, the verbs of pvc's, the predicates of certain idioms and certain predicates with resultative xcomps belong here. These types also have the [+VM] feature. In addition, there are predicates without the [+VM] feature that also need [-VM]. For example, nominal and adjectival predicates, and verbs that always need word-initial stress belong here. On partitive mass and plural NPs in Estonian, corresponding to bare nominal VMs in Hungarian, see Tamm (2007a,c).

<sup>25</sup>In her review, Anne Tamm points out that there are similarities between Laczkó's analysis of Hungarian particle verbs and the analysis of Estonian particle verbs and aspect in Rätsep (1969) written in Estonian, which Tamm (2012c: 62–63, 72–75) has summarized, or Rätsep's (1978) account of government structures of complex verbs in Estonian.

Verbs that combine with *ára* in this use have a theme or patient argument, obligatorily realized as a subject or an object. The closest Hungarian equivalents are *meg* ‘PFV’ (as in *meg-hal* ‘PFV-die’) and *el* ‘away’ (as in *el-olvad* ‘away-become.melted’).

- (11) Bounding use of *ära*, Estonian:  
Ta suudles tüdruku ära.  
s/he kiss.PST.3SG girl.GEN ÄRA  
'S/he did the kissing of a girl.'

This sentence is appropriate in the following situation, for instance. Someone makes a bet to kiss a girl, and when this goal is achieved, the result can be reported by using this PVC. The closest Hungarian counterparts are *meg* ‘PFC’ (*meg-ebédel* ‘have/eat up one’s lunch’) and *ki* ‘out’ (as in *ki-alussza magát* ‘out-sleep oneself.ACC’ ‘sleep one’s share, as much as needed’).

Tamm assumes that *ära* in its directional use has a PRED feature, and she gives the following lexical representation (Tamm 2004c: 231).

- (12) *ära* P  $(\uparrow \text{PRED}) = \text{'AWAY}\langle(\uparrow \text{SUBJ})\rangle\text{'}$   
 $\{((\text{XCOMP } \uparrow) \text{ B1}) \vee ((\text{XCOMP } \uparrow) \text{ B2})\}$

This encodes that the particle functions as the **PRED** of the lexical verb, and it has a subject argument. In addition, it has disjunctive existential constraints on the boundedness (**B**) attributes.

Tamm assumes that *ära* in its completive use also has a PRED feature, see her lexical form in (13) (Tamm 2004c: 232).

- (13) *ära* P ( $\uparrow$  PRED) = 'UP, COMPLETELY( $\langle \uparrow$  SUBJ  $\rangle$ )'  
 $\{ ((\text{XCOMP } \uparrow) \text{ B1}) = \text{MAX} \vee ((\text{XCOMP } \uparrow) \text{ B2}) \}$

As opposed to its previous two uses, Tamm assumes that *ära* in its bounding use has no PRED feature, and it only encodes B and focus specifications, see (14) (Tamm 2004c: 229).

- (14) ära                  Prt    (↑ B1) = MAX  
                                (↑ B2) = MIN  
                                (↑ FOCUS B1) = MAX  
                                (↑ FOCUS B2) = MIN

The particle in this use contributes f-structure information about the aspectual features of the clause, see the first two annotations, and it also encodes that this boundedness is the focused information, see the last two annotations.

*Tibor Laczkó*

In addition, verbal predicates can also carry aspectual information in their lexical forms. For instance, Tamm assumes that *suuddlema* ‘kiss’, see (11) for instance, has the following lexical representation.

- (15) *suuddlema* V (↑ PRED) = ‘KISS<(↑ SUBJ)(↑ OBJ)>’  
(↑ B2)

This verb has an existential constraint on B2, which can be unified with the MIN value of the B2 of the particle in (14). Finally, the partitive and total case-markers on object arguments also encode aspectual information, so the entire aspectual feature value set of an Estonian sentence comes from three main sources via unification: verbs, aspectual particles and partitive/total case markers.<sup>26</sup>

### 3.3 Concluding remarks

At the end of Section 3 we can make the following concluding remarks.

Hungarian VM phenomena are relevant from both cross-theoretical and LFG-specific perspectives in two important respects.

First, the focus-VM complementarity is a general generative theoretical issue. As the foregoing discussion shows, LFG provides a flexible formal platform even for alternative analyses significantly different in nature, which may be due to partially different views of the relevant components of the architecture of LFG.

Second, the behaviour of Hungarian PVCs, representing the major class of VMs, is of great importance in the realm of complex predicates across typologically different languages, see Alsina et al. (1997) in general and Ackerman & Lesourd (1997) in that volume, in particular. The mixed lexical-morphological and syntactic properties of compositional and productive as well as non-compositional and unproductive PVCs pose a substantial challenge for both syntactically and lexically oriented generative theories, including LFG. From their entirely lexicalist perspective, Ackerman et al. (2011) give a taxonomic overview of a variety of approaches to complex predicates in LFG and HPSG. They point out that the classical models of the two theories rejected argument-structure-changing operations in the syntax, including complex predicate formation: see Bresnan (1982) and Pollard & Sag (1987). However, some more recent views in both theories admit syntactic complex predicate formation: see Alsina (1992, 1997), Butt (2003) and Müller (2006). By contrast, Ackerman et al. (2011), in their Realization-Based Lexicalism (RBL) model, reject complex predicate formation in the syntax, and,

<sup>26</sup>On the aspectual interaction of various verb types and partitive/total case in Estonian, see the discussion of Tamm’s (2006) analyses in Section 6.2.



as a trade-off, they admit analytic, i.e. multiple-word, forms of predicates in their lexicon as a marked option. As regards the treatment of Hungarian pvc's, Ackerman (2003) develops an RBL analysis. Forst et al. (2010), Laczkó & Rákosi (2011), Rákosi & Laczkó (2011) and Laczkó & Rákosi (2013), in their LFG-XLE framework, handle the productive types in the syntax by means of the *RESTRICTION* operator. By contrast, Laczkó (2013a), in the same framework, argues that both productive and unproductive pvc's need a lexical treatment.

As regards Estonian, Tamm's (2004c) analysis has demonstrated that LFG also provides an appropriate formal apparatus for capturing the interplay of discourse functionality and the complex, multidimensional aspectuality system of this language.

## 4 Negation in Hungarian

Miestamo et al. (2015) discuss negation in Uralic in a comprehensive and systematically comparative fashion.<sup>27</sup> They show that 17 Uralic languages employ negative auxiliaries. Hungarian, Khanty, Mansi and Estonian are exceptions in that they have no such auxiliaries. Of all these languages, we are only aware of a few LFG analyses of negation in Hungarian (most of them being rather sketchy and covering only some aspects of negation phenomena).

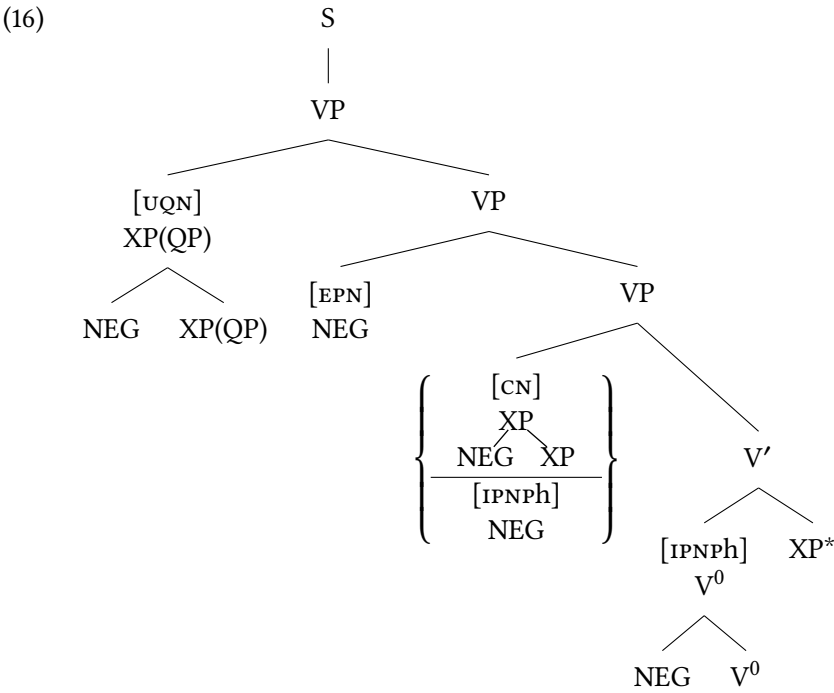
Laczkó (2014c) develops the first comprehensive LFG-XLE approach to the following six major types of clausal (aka predicate) and constituent negation in Hungarian: (i) ordinary constituent negation (the negated constituent is focused); (ii) universal quantifier negation without (another) focused element (= ordinary constituent negation, i.e. the negated universal quantifier is focused); (iii) universal quantifier negation with focus (= there is a preverbal focused constituent following the negated universal quantifier); (iv) predicate negation, without focus, the negative particle precedes the verb; (v) predicate negation, with focus, the negative particle precedes the verb; (vi) predicate negation, with focus, the negative particle precedes the focus.<sup>28</sup> He proposes the following structural analysis.<sup>29</sup>

<sup>27</sup>There is a publicly accessible database on negation in Ob-Ugric and Samoyedic languages at <https://www.univie.ac.at/negation/index-en.html>.

<sup>28</sup>Payne & Chisarik (2000), in their OT-LFG framework, also sketch an analysis of some of these types. For a critical overview, see Laczkó (2014c).

<sup>29</sup>In (16) NEG stands for the (category of the) negative particle and the abbreviations in square brackets indicate the types of negation: [uQN] = universal quantifier negation, [EPN] = (VP-)external predicate negation, [CN] = constituent negation, [IPNPh] = (VP-)internal predicate negation, phrasal adjunction, [IPNH] = (VP-)internal predicate negation, head-adjunction. The curly brackets signal the complementarity of [CN] and [IPNPh].

Tibor Laczkó



In XLE grammars three devices are used for the encoding of negation: (i) the negative morpheme (whether bound or free) can be represented as a member of the ADJUNCT set; (ii) it can encode the  $[\text{NEG} +]$  feature value; (iii) it can encode the  $[\text{POL NEG}]$  feature value. Laczkó (2015) points out that these devices are not used uniformly or consistently across the XLE grammars of various languages. He makes the following proposal. Type (i) is most appropriate when a language uses a free morpheme for the expression of negation, a negative particle. Type (ii) is best for bound negative morphemes. Type (iii) is most natural for encoding the scope of negation. In this proposed system, he develops an LFG-XLE analysis of Hungarian negative concord items.

In Laczkó's (2014c) approach negated constituents also occupy the Spec,VP position, see  $[\text{CN}]$  in (16). In addition, in the case of clause negation, the negative particle is also assumed to be in Spec,VP when there is no focused constituent there, see  $[\text{IPNPh}]$ . In his rules, Laczkó assumes that the negative particle has the category NEG, and he uses a special XLE-style phrasal categorial label for negated constituents: XPneg. His XPneg rule is given in (17).

$$(17) \quad \text{XPneg} \longrightarrow \begin{array}{cc} \text{NEG} & \text{XP} \\ \downarrow \in (\uparrow \text{ADJ}) & \end{array}$$

On the basis of these assumptions and rules, he adds the following two disjuncts to the disjunction in Spec,VP established so far, handling focused constituents and vms, shown in (8) in Section 3.1.<sup>30</sup>

- (18) { ... | XP<sub>neg</sub> | NEG }
- |             |             |
|-------------|-------------|
| (↑ GF)=↓    | ↓ ∈ (↑ ADJ) |
| (↑ FOCUS)=↓ | (↑ FOCUS)=↓ |

As this section has shown, LFG provides an inventory of appropriate formal devices for analyzing complex negation phenomena in languages like Hungarian. At the same time, the treatment of these negation phenomena motivates examining the nature of the relevant formal devices carefully.

## 5 Copula constructions

### 5.1 Hungarian

The two major general LFG strategies for the treatment of copula constructions (ccs) across languages are represented by Butt et al. (1999) and Dalrymple et al. (2004). In the former approach, ccs are treated in a uniform manner functionally. The copula is always assumed to be a two-place predicate. It subcategorizes for a subject (SUBJ) argument, which is uncontroversial in any analysis of these constructions, and the other constituent is invariably assigned a special, designated function designed for the second, ‘postcopular’ argument of the predicate: PREDLINK. As opposed to this approach, in Dalrymple et al.’s (2004) view, the SUBJ & PREDLINK version is just one of the theoretically available options. In addition, they postulate that the copula can be devoid of a PRED feature (and, consequently, argument structure) and in this use it only serves as a pure carrier of formal verbal features: tense and agreement. Finally, it can also be used as a one-place ‘raising’ predicate, associating the xCOMP function with its propositional argument and also assigning a non-thematic SUBJ function. When the postcopular constituent has the PREDLINK function, it is closed in the sense that its subject argument is never realized outside this constituent. The xCOMP and the PREDLINK types involve two semantic and functional levels (tiers): the copula selects the relevant constituent as an argument. By contrast, when the copula is a mere formative, the two elements are at the same level (tier): the postcopular constituent is the real predicate and the copula only contributes morpho-syntactic features.

<sup>30</sup>Based on their prosodic and semantic behaviour, he assumes that both types of negative elements are focused constituents.

Tibor Laczkó

In LFG’s formal system, they are functional coheads. All this is summarized in Table 2.

Table 2: Three types of copular constructions, Dalrymple et al. (2004)

role of the postcopular constituent		
open		closed
(A)	(B)	(C)
main PRED, the copula is a formative: functional coheads (single-tier)	XCOMP of the copula main PRED: 'be<(↑ XCOMP)>(↑ SUBJ)' (double-tier)	PREDLINK of the copula main PRED: 'be<(↑ SUBJ)>(↑ PREDLINK)' (double-tier)

As regards the treatment of copula constructions, Laczkó (2012) develops the first comprehensive LFG analysis of the following five most important types of ccs in Hungarian: (i) attribution or classification; (ii) identity; (iii) location; (iv) existence; (v) possession. He subscribes to the view, advocated by Dalrymple et al. (2004) and also by Nordlinger & Sadler (2007), among others, that the best LFG strategy is to examine all ccs individually, and to allow for diversity and systematic variation both in c-structure and in f-structure representations across and even within languages. This means that he rejects Butt et al.’s (1999) and Attia’s (2008) uniform PREDLINK approach at the f-structure level. Table 3 summarizes the most important aspects of his analysis.<sup>31</sup>

Here we can only highlight the most crucial ingredients of this approach, concentrating on the ‘copula’s function’ and ‘argument structure’ columns in the table. In the attribution/classification type the copula has no PRED feature and, thus, no argument structure, cf. column (A) in Table 2. The versions of the copula in all the other four cc types are two-place predicates. In the identity and possession types the second argument is assumed to have the PREDLINK function, cf. column (C) in Table 2, while in the location and existence ccs it bears the OBL<sub>LOC</sub> function, which is a variant of the closed type of postcopular constituents in

<sup>31</sup>The following abbreviations are used in Table 3: COP = copula, ATTR/CLASS = attribution/classification, PR3: COP = is the copula present in the present tense and 3<sup>rd</sup> person paradigmatic slots? PR3: NEG = how is negation expressed in PR3? VM = which element (if any) occupies the VM position in neutral sentences? S = SUBJ, PL = PREDLINK, interch = the two arguments’ grammatical functions are interchangeable in the 3<sup>rd</sup> person, spec = specific, def = definite, FOC = FOCUS, agr = agreement.

Table 3: Laczkó’s (2012) analysis of Hungarian copula constructions

CC TYPE	PR3		COPULA’S FUNCTION	ARGUMENT STRUCTURE	OTHER VM	OTHER TRAITS
	COP	NEG				
ATTR/CLASS	-	<i>nem</i>	formative	-	AP/NP	NP: -spec
IDENTITY	-	<i>nem</i>	predicate	<S, PL>	SUBJ	S: +spec, interch.
LOCATION	+	<i>nincs</i>	predicate	<S, OBL>	OBL	S: +spec
EXISTENCE	+	<i>nincs</i>	predicate	<S, (OBL)>	-	S: -spec COP: FOC
POSSESSION	+	<i>nincs</i>	predicate	<S, PL>	-	S: -def S&PL agr. COP: FOC

column (C) in Table 2. Thus, in Laczkó’s (2012) analysis the copula has five distinct lexical forms, which encode their respective sets of properties indicated in Table 3.

5.2 Inari Saami and Finnish

Toivonen (2007) analyzes subject-verb agreement phenomena in Inari Saami with a brief comparison with the corresponding Finnish phenomena, see Section 7.1.1. In her general approach, she also proposes an LMT (Lexical Mapping Theory: **chapters/Mapping**) analysis of Inari Saami possessive constructions, again with a brief comparison with the Finnish counterparts. The empirical generalizations that she starts with, and which are relevant here, are as follows. (i) The possessed item is the subject. (ii) The possessed item bears nominative case. (iii) The possessor bears locative case. Consider one of her examples in (19), illustrating these facts.

- (19) Inari Saami:  
Muste lah tun.  
I.LOC are.2SG you.NOM.SG  
‘I have you.’

Toivonen assumes that the Inari Saami copula in this function is a two-place predicate with a theme (possessum) argument and a location (possessor) argument

*Tibor Laczkó*

that receive the [-r] and the [-o] intrinsic specifications, respectively, and they are mapped onto SUBJ and OBL, respectively: see (20).

(20)

		theme	location	
<i>leðe</i>	<	x	y	>
		[-r]	[-o]	
		SUBJ	OBL	

Toivonen compares Inari Saami and Finnish possessive constructions. For her comparison from the perspective of agreement, see Section 7.1.1. Here we concentrate on the GFS of the arguments of the possessive copulas of the two languages. Compare Toivonen's Inari Saami example in (19) above with her corresponding Finnish example in (21).

(21) Finnish:

Minulla	on	sinut.
I.ADE	is.3SG	you.ACC.SG
'I have you.'		

She makes the following generalizations about Finnish possession ccs. The possum is either in nominative case (ordinary noun phrases) or in accusative case, see (21), and it has the OBJ function. The possessor is an oblique case-marked noun phrase, and it has the SUBJ function.

These two sections have shown that the behaviour of copula constructions in Hungarian, Inari Saami and Finnish exhibits remarkable variation, especially in the case of possession ccs. We can make the following concluding observations. On the one hand, the LFG framework, in this case, too, provides appropriate formal tools for feasible analyses of these construction types. On the other hand, the complexity of these phenomena can be used to argue for particular approaches in the inventory of LFG's alternative formal devices in this particular domain.

## 6 Aspects of argument realization

### 6.1 Finnish

Pylkkänen (1997) develops an event-structure-based linking approach to Finnish causatives. She claims that her theory is minimalistic in two respects. On the one hand, in formalizing the relationship between event participants it minimizes reference to the thematic role properties of these participants (e.g. agent, theme and

6 *LFG and Finno-Ugric languages*

cause) by referring to events themselves. The basic assumption is that if one eventuality causes another then the participants of the former always rank higher than those of the latter. On the other hand, an adequately developed system of inferring prominence relations obviates the need for argument structure, the level of representation mediating between event structure and grammatical functions. Pylkkänen's system of inferring prominence from lexical semantic representations capitalizes on the following two assumptions proposed by Parsons (1990): (i) thematic roles are relations between events and individuals; (ii) causation is a relation between events. As a consequence, the thematic hierarchy is treated as applying at the level of individual events and not at the level of predicates. From this it follows that a predicate can have more than one thematic hierarchy: as many thematic hierarchies as events. All participants can be organized into a prominence hierarchy by ranking the individual thematic hierarchies with respect to each other. This ranking is regulated by Parsons' second assumption: the causal relations between events. In essence, if  $E1$  CAUSES  $E2$ , then  $E1_{\Theta H}$  (the thematic hierarchy of  $E1$ ) is ranked higher than  $E2_{\Theta H}$  (the thematic hierarchy of  $E2$ ). Consider Pylkkänen's two hierarchies in (22) and (23).

(22) Thematic Hierarchy:      agent/experiencer > other > theme

(23) Event Hierarchy:       $\text{cause}(e1, e2) \rightarrow E1_{\Theta H} > E2_{\Theta H}$

Then linking constraints provide the mapping between the prominence hierarchy resulting from (22) and (23) and the following grammatical function hierarchy.

(24) SUBJ > OBJ > OBJ <sub>$\theta$</sub>  > OBL

In order for the linking constraints to be unifiable, Pylkkänen converts Parsons' logical forms into attribute-value matrices. Consider her f-structure and event structure representation of (25), one of her examples, in (26).

(25) Finnish:

Matti      kävel-yttä-ä      koiraa.  
 Matti.NOM walk-CAUS-3SG dog.PAR  
 'Matti walks the dog.'

In the event structure there are ranked participants. IND means 'index', which is a 'pointer' to an event participant, and RANK indicates the prominence of the participant concerned. In the case of (25), the RANK1 participant is realized as the subject, while the RANK2 and RANK3 participants are realized as the object.

*Tibor Laczkó*

$$(26) \quad \left[ \begin{array}{c} \text{F-STR} \\ \\ \text{EVENTSTR} \end{array} \left[ \begin{array}{c} \text{SUBJ} \left[ \begin{array}{c} \text{PRED MATTI} \\ \text{CASE NOM} \end{array} \right] \\ \text{OBJ} \left[ \begin{array}{c} \text{PRED DOG} \\ \text{CASE PAR} \end{array} \right] \\ \text{E1} \left[ \begin{array}{c} \theta\_RELS \\ \text{AGENT} \left[ \begin{array}{c} \text{IND MATTI} \\ \text{RANK 1} \end{array} \right] \\ \text{THEME} \left[ \begin{array}{c} \text{IND DOG} \\ \text{RANK 3} \end{array} \right] \end{array} \right] \\ \text{E2} \left[ \begin{array}{c} \theta\_RELS \\ \text{AGENT} \left[ \begin{array}{c} \text{IND DOG} \\ \text{RANK 3} \end{array} \right] \\ \text{SEM\_TYPE WALK} \end{array} \right] \\ \text{REL CAUSE(E1,E2)} \end{array} \right] \right]$$

## 6.2 Estonian

Tamm (2006) develops an LFG analysis of the interaction of transitive telic verbs and aspectual case in Estonian. In this language the objects of telic verbs can bear either partitive (PAR) case or total (TOT) case. The choice between partitive and total is regulated by the aspectual features of the sentence, compare Tamm’s examples in (27) and (28).<sup>32</sup>

- (27) Estonian:  
 Mari kirjutas raamatu ühe aastaga.  
 Mari.NOM write.PST.3SG book.TOT one.GEN year.COM  
 ‘Mari wrote a/the book in a year.’

- (28) Estonian:  
 Mari kirjutas raamatut terve aasta.  
 Mari.NOM write.PST.3SG book.PAR whole.TOT year.TOT  
 ‘Mari was writing a/the book for a whole year.’

Tamm shows that the sentence in (27), with its object in total case, has a perfective interpretation, and the sentence in (28), with its object in partitive case, is imperfective, as is supported by the types of the adjuncts in them: ‘in a year’ vs. ‘for a year’. In addition, Tamm shows that Vendlerian achievement verbs like

<sup>32</sup>The lexical entries for the Estonian case-markers encoding aspectual features are modeled as semantic (Butt & King 2004) and constructive cases (Nordlinger & Sadler 2004), and they provide the formal tools for Tamm’s analysis. On the terminology of Finnic core cases, see Tamm (2011a, 2012c). On partitives in Finnish, see Vainikka & Maling (1996).



## 6 LFG and Finno-Ugric languages

*võitma* ‘defeat’ are compatible with objects in partitive case in Estonian, although the sentences they occur in are perfective by default, see her example in (29).

- (29) Estonian:  
 Mari        *võitis*                Jürit.  
 Mari.NOM defeat.PST.3SG George.PAR  
 ‘Mary defeated George.’

In her analysis, Tamm introduces the boundedness aspectual feature: *B* with two values: *MIN* and *MAX*. She associates this feature both with the lexical forms of the two transitive verb types seen above and with the lexical representations of case markers in the following way. Her basic generalization is that ‘write’-type verbs are boundable, and ‘defeat’-type verbs are bounded. In the lexical form of the former boundedness is encoded as an existential constraint, while in the lexical form of the latter it is encoded as a defining equation: the *B* feature has the *MIN* value, see (30) and (31), respectively.

- (30) *kirjutama* ‘WRITE’...        ( $\uparrow B$ )  
 (31) *võitma* ‘DEFEAT’...        ( $\uparrow B$ ) = *MIN*

As regards case, the total case-marker, attached to an object noun phrase, introduces the *MAX* value for *B*, while the partitive case-marker specifies *B* as  $\neq \text{MAX}$ . These values are encoded with inside-out function application, see (32) and (33).

- (32) *tot*                ( $\uparrow \text{CASE}$ ) = *TOT*  
                           ( $((\text{OBJ } \uparrow) B)$ ) = *MAX*  
 (33) *par*                ( $\uparrow \text{CASE}$ ) = *PAR*  
                           ( $((\text{OBJ } \uparrow) B)$ )  $\neq \text{MAX}$

In this system, a ‘write’-type verb requires that the sentence should be marked for boundedness, and its underspecified *B* feature admits either of the two object cases. For instance, Tamm gives the following lexical representations for the verb and the object in (27).

- (34) *kirjutas*        *V*        ( $\uparrow \text{PRED}$ ) = ‘WRITE $\langle(\uparrow \text{SUBJ})(\uparrow \text{OBJ})\rangle$ ’  
   ( $\uparrow \text{TENSE}$ ) = *PST*  
   ( $\uparrow \text{PERS}$ ) = 3  
   ( $\uparrow \text{NUM}$ ) = *SG*  
   ( $\uparrow B$ )

*Tibor Laczkó*

- (35) *raamatu*    N    ( $\uparrow$  PRED) = 'BOOK'  
                               ( $\uparrow$  CASE) = TOT  
                               ((OBJ  $\uparrow$ ) B) = MAX

On the basis of this, her f-structure representation of (27) is as follows.

- (36) 
$$\left[ \begin{array}{ll} \text{PRED} & \text{'WRITE<SUBJ, OBJ>'} \\ \text{B} & \text{MAX} \\ \text{TNS} & \text{PST} \\ \text{NUM} & \text{SG} \\ \text{PERS} & 3 \\ \text{SUBJ} & \left[ \begin{array}{l} \text{PRED 'MARI'} \\ \text{CASE NOM} \\ \dots \end{array} \right] \\ \text{OBJ} & \left[ \begin{array}{l} \text{PRED 'BOOK'} \\ \text{CASE TOT} \\ \text{NUM SG} \\ \dots \end{array} \right] \end{array} \right]$$

Obviously, the f-structure representation of (28) would be different from (36) in one important respect: the value of B would be  $\neq$ MAX on the basis of (33).

By contrast, the value of the B feature of a 'defeat'-type verb is MIN, which only allows compatibility with an object in partitive case, given that total case encodes the opposite value: MAX. Tamm offers the following lexical representations for the verb and the object in (29), and she points out that there is no value clash with respect to the B feature.

- (37) *võitis*        V    ( $\uparrow$  PRED) = 'DEFEAT<( $\uparrow$  SUBJ) ( $\uparrow$  OBJ)>'  
                               ( $\uparrow$  TNS) = PST  
                               ( $\uparrow$  PERS) = 3  
                               ( $\uparrow$  NUM) = SG  
                               ( $\uparrow$  B) = MIN
- (38) *Jürit*        N    ( $\uparrow$  PRED) = 'GEORGE'  
                               ( $\uparrow$  CASE) = PAR  
                               ((OBJ  $\uparrow$ ) B)  $\neq$  MAX

As another argument-realization topic, Torn (2006) discusses the status of certain non-core arguments and adjuncts of verbal predicates in Estonian. She points out that fundamentally there are two approaches to these constituents. One of them regards non-core arguments as oblique case-marked indirect objects,

## 6 LFG and Finno-Ugric languages

separating them from adjuncts, while the other lumps the two groups together as adverbials. Torn subscribes to the first approach.

By way of illustration, Torn shows that in this language participants of an event that are indirectly affected are realized by noun phrases bearing the same ‘local’ case suffixes as are used to express spatial adverbial dependents: see her examples in (39) and (40).

- (39) Adverbial allative, Estonian:

Mees istus diivanile.  
man.NOM sat sofa.ALL  
‘A man sat onto the sofa.’

- (40) Oblique allative, Estonian:

Ema andis lapsele raha.  
mother.NOM gave child.ALL money.PAR  
‘The mother gave money to the child.’

Torn says that *diivanile* ‘onto the sofa’, a noun phrase in allative case, is an un-governed adverbial constituent in (39), while *lapsele* ‘to the child’, a noun phrase in allative case here, too, expresses the indirectly affected argument of the ditransitive verb *andma* ‘give’ in (40). In her terminology, *diivanile* in (39) is an adverbial modifier, and *lapsele* in (40) is an object adverbial.

Torn offers the following three arguments for distinguishing object adverbials from adverbial modifiers. (i) A verbal predicate selects a particular governed case for its object adverbial and not a semantically compatible set of cases. (ii) An object adverbial constituent can serve as an antecedent in an obligatory control construction. (iii) It is a functional similarity between object adverbials on the one hand, and subjects and objects on the other, that they can be involved in systematic case alternations. Such alternations can never involve adverbial modifiers.

Torn adopts LFG’s LMT classification of governable grammatical functions. In this setting, she assumes that locative case-marked noun phrases can have either the OBL or the ADJUNCT function.

Tibor Laczkó

## 7 Morpho-syntactic phenomena

### 7.1 Agreement

#### 7.1.1 Subject-verb agreement in Inari Saami and Finnish

Toivonen (2007) examines verbal inflectional morphology in Inari Saami. She develops her analysis by concentrating on the Saami copula *lede* ‘to be’. In this language, as in various Northern Uralic languages, the number feature has three possible values: singular, dual and plural. It is another special property of this language that there can be either full agreement or partial agreement between the subject and the verb. Animate and specific subjects trigger the former, see (41), inanimate subjects trigger the latter, see (42).

- (41) Inari Saami:  
 Meecest lava uábbi já viljá.  
 forest.LOC are.3DU sister.NOM and brother.NOM  
 ‘In the forest are my sister and brother.’

- (42) Inari Saami:  
 Riddoost láá kyehti keeđgi.  
 beach.LOC are.3PL two rock  
 ‘On the beach are two rocks.’

Subject noun phrases headed by unspecific human nouns and animal nouns can trigger either full or partial agreement. (43) illustrates the unspecific human case.

- (43) Inari Saami:  
 Táálust lava/láá kyehti ulmuu.  
 house.LOC are.3DU/are.3PL two person  
 ‘There are two people in the house.’

Toivonen presents the paradigms of the copula in this three-way number and dual agreement system as in Table 4. She develops an LFG analysis with fully specified and underspecified lexical forms of verbal predicates. Consider her representations of four morphological forms of the copula in (44–47).

- (44) *lava*            V    (↑ PRED) = ‘BE’  
                               (↑ TENSE) = PRS  
                               (↑ MOOD) = INDICATIVE  
                               (↑ SUBJ NUM) = DU  
                               (↑ SUBJ PERS) = 3  
                               (↑ SUBJ HUM) = +

Table 4: Agreement paradigms for ‘to be’

		full	partial
SG	1	lam	lii
	2	lah	lii
	3	lii	lii
DU	1	láán	láá
	2	leppee	láá
	3	lava	láá
PL	1	lep	láá
	2	leppeδ	láá
	3	láá	láá

- (45) *lam*            V    (↑ PRED) = ‘BE’  
                          (↑ TENSE) = PRS  
                          (↑ MOOD) = INDICATIVE  
                          (↑ SUBJ NUM) = SG  
                          (↑ SUBJ PERS) = 1  
                          (↑ SUBJ HUM) = +

- (46) *lii*            V    (↑ PRED) = ‘BE’  
                          (↑ TENSE) = PRS  
                          (↑ MOOD) = INDICATIVE  
                          (↑ SUBJ NUM) = SG

- (47) *láá*            V    (↑ PRED) = ‘BE’  
                          (↑ TENSE) = PRS  
                          (↑ MOOD) = INDICATIVE

Toivonen makes crucial use of the principle of morphological blocking as developed by Andrews (1990). The basic idea is that if a subject noun phrase is compatible with more than one verb form, it will select the variant that exhibits the largest number of its own feature values. This explains, for instance, why human subjects do not freely co-occur with *láá* or why singular subjects cannot co-occur with *láá*. The answer to the first question is that *láá* has no [+human] feature, see (47). The answer to the second question is that there are more specific forms of the copula in that they also encode the [+singular] feature, compare (47) with (45) and (46).

*Tibor Laczkó*

Toivonen also briefly compares the Inari Saami agreement system with the corresponding Finnish system. She points out that Finnish has no grammatical dual. In addition, Finnish does not exhibit partial agreement. Furthermore, animacy has not been grammaticalized in standard Finnish. It is another significant difference that in Inari Saami, verb agreement is always triggered by grammatical subjects, while in Finnish several independent conditions need to be simultaneously satisfied for agreement to take place. First, in Finnish, as well as in Estonian,<sup>33</sup> only nominative NPs trigger agreement, compare Toivonen's examples in (48) and (49).

(48) Finnish:

Autot ajavat yleensä kovaa moottoriteillä.  
cars.NOM drive.3PL generally hard motorways.ADE  
'Cars generally drive fast on the motorways.'

(49) Finnish:

Linja-autoja kulkee nykyisin joka sunnuntai.  
buses.PAR run.3SG nowadays every Sunday  
'Nowadays, buses run every Sunday.'

In (48) the nominative subject triggers agreement, while in (49) the subject is in partitive case and the verb takes 3SG default agreement.

A Finnish verb also has default agreement in existential and possessive constructions. (50) illustrates the latter type.

(50) Finnish:

Koulussa on uudet opettajat.  
school.INE is.3SG new.NOM.PL teachers.NOM  
'The school has new teachers.'

In this example, although the (post-verbal) subject is nominative, it is not in its preverbal canonical position; therefore, here, too, the verb displays 3SG default agreement.

As regards their possessive constructions, Inari Saami and Finnish differ in two significant respects. On the one hand, in Inari Saami possessive constructions pronouns are in nominative case, while in Finnish the corresponding pronouns take accusative case, compare (51) and (52). On the other hand, the possessum is always in nominative case in Inari Saami, it has the subject function, and it always triggers agreement, while in Finnish the possessum is either in nominative

---

<sup>33</sup>See Hiietam (2003), for instance.

6 *LFG and Finno-Ugric languages*

case (ordinary noun phrases) or in accusative case, and the verb always carries 3SG default agreement, compare (53) and (54).

- (51) Inari Saami:  
 Muste lah tun.  
 I.LOC are.2SG you.NOM.SG  
 'I have you.'
- (52) Finnish:  
 Minulla on / \*olen sinut.  
 I.ADE is.3SG / is.1SG you.ACC.SG  
 'I have you.'
- (53) Inari Saami:  
 Muste lava puásui já peenuv.  
 I.LOC are.3DU reindeer.NOM and dog.NOM  
 'I have a reindeer and a dog.'
- (54) Finnish:  
 Minulla on / \*olen poro ja koira.  
 I.ADE is.3SG / is.1SG reindeer.NOM and dog.NOM  
 'I have a reindeer and a dog.'

Toivonen makes the following concluding generalization about Finnish possessive constructions. There is no normal agreement in them, because the posses-sum is not the subject, and because the subject possessor is not in nominative case. This is why 3SG default agreement is employed.

Toivonen offers a comparative overview of the agreement systems of Inari Saami and Finnish shown in Table 5.

### 7.1.2 Aspects of differential object marking in Uralic

Coppock & Wechsler (2010) point out that there is object agreement in Nenets, Enets, Nganasan and Selkup in the Samoyedic family and in Mordvinian (Finno-Volgaic), Hungarian (Ugric), Ostyak and Vogul (both Ob-Ugric) in the Finno-Ugric family. These languages exhibit remarkable variation with respect to the feature specifications of their object agreement. In Hungarian and Samoyedic there are two conjugation paradigms: subjective and objective, and the latter is used in the case of definite and third person objects. In Ob-Ugric languages there is a subjective conjugation and three objective conjugation paradigms, one

Table 5: Agreement in Inari Saami and Finnish

	Inari Saami	Finnish
Partial agreement	☒	
Default agreement		☒
Animacy effects	☒	
Agreement in possessive construction	☒	
Agreement in existential construction	☒	
Possessed nouns in nominative case	☒	☒
Possessed pronouns in nominative case	☒	

for each possible number value of the object (singular, dual and plural). In Mordvinian there is genuine agreement for both person and number between the verb and the object. Coppock & Wechsler (2010) concentrate on Northern and Eastern Ostyak, Hungarian and Samoyedic.<sup>34</sup>

In Northern Ostyak the verb agrees with its object in number but not in person: see (55) and (56). An additional factor is that the object has to be topical, otherwise the subjective conjugation is used.

- (55) Northern Ostyak:  
Ma tām kālang wel-sə-l-am.  
I this reindeer kill-PST-PLOBJ-1SGSUBJ  
‘I killed these reindeer.’

- (56) Northern Ostyak:  
Xūnsi nāng mūng-iluw xālsa want-lə-l-an?  
when you we-ACC where see-PRS-PLOBJ-2SGSUBJ  
‘When did you see us where?’

Coppock & Wechsler (2010) postulate the following diachronic analysis of these facts.

At the first stage third person pronouns were incorporated ( $\downarrow$  PRED)=‘PRO’ and ( $\downarrow$  INDEX PERS)=3 with the three number values ( $\downarrow$  INDEX NUM)=*n*. This was combined with the topicality condition: ( $\downarrow_{\sigma}$  DF)=TOPIC.

<sup>34</sup>Also see Coppock & Wechsler (2012) on Hungarian.



## 6 LFG and Finno-Ugric languages

- (57)  $V_{aff}$        $(\uparrow \text{OBJ}) = \downarrow$   
                   $(\downarrow \text{PRED}) = \text{'PRO'}$   
                   $(\downarrow_{\sigma} \text{DF}) = \text{TOPIC}$   
                   $(\downarrow \text{INDEX PERS}) = 3$   
                   $(\downarrow \text{INDEX NUM}) = N$     where  $N \in \{\text{SG, DU, PL}\}$

At the second stage the PRED 'pro' was dropped.

- (58)  $V_{aff}$        $(\uparrow \text{OBJ}) = \downarrow$   
                   ~~$(\downarrow \text{PRED}) = \text{'PRO'}$~~   
                   $(\downarrow_{\sigma} \text{DF}) = \text{TOPIC}$   
                   $(\downarrow \text{INDEX PERS}) = 3$   
                   $(\downarrow \text{INDEX NUM}) = N$     where  $N \in \{\text{SG, DU, PL}\}$

The authors claim that it is reasonable to assume that at this stage person specification was present because Eastern Ostyak still manifests this stage.

At the third stage the person specification was lost in Northern Ostyak, see (59), but this did not happen in Eastern Ostyak.

- (59)  $V_{aff}$        $(\uparrow \text{OBJ}) = \downarrow$   
                   ~~$(\downarrow \text{PRED}) = \text{'PRO'}$~~   
                   $(\downarrow_{\sigma} \text{DF}) = \text{TOPIC}$   
                   ~~$(\downarrow \text{INDEX PERS}) = 3$~~   
                   $(\downarrow \text{INDEX NUM}) = N$     where  $N \in \{\text{SG, DU, PL}\}$

As a result, these objective conjugation suffixes became usable with first and second person objects, too.

Coppock & Wechsler (2010) also show that Hungarian has two conjugations that are conditioned by the definiteness of the object by using the following examples. The general pattern is that definite objects trigger the objective agreement type, see (60), and indefinite objects require the subjective type, see (61).

- (60) Hungarian:  
       Lát-om            a    madar-at.  
       see-PRS.1SG.DEF the bird-ACC  
       'I see the bird.'

- (61) Hungarian:  
       Lát-ok            egy madar-at.  
       see-PRS.1SG.INDF a    bird-ACC  
       'I see a bird.'

*Tibor Laczkó*

In addition, the objective agreement type is sensitive to the person feature of the object: in the pronominal domain only third person pronouns trigger it, see (62), while first and second person pronouns require the subjective conjugation, see (63). Coppock & Wechsler (2010) refer to this as the third person restriction in this language.

- (62) Hungarian:  
 Lát-ják                      őt/őket.  
 see-PRS.3PL.DEF it/them  
 ‘They see it/them.’

- (63) Hungarian:  
 Lát-nak                      engem/téged/minket.  
 see-PRS.3PL.INDF me/you/us  
 ‘They see me/you/us.’

It is another property of the Hungarian object agreement system that it is not sensitive to the number value of the object.

Coppock & Wechsler (2010) propose the following diachronic analysis. At the first stage, just like in the case of Northern and Eastern Ostyak, third person pronoun incorporation took place, see (57) above. The second stage was also the same: the PRED ‘pro’ was dropped and the topicality condition retained, see (58). This is the present-day Eastern Ostyak system. At the third stage the number constraint was dropped, but the person restriction was retained, see (64) and compare it with (59) characterizing Northern Ostyak.

- (64)  $V_{aff}$                        $(\uparrow \text{OBJ}) = \downarrow$   
                                       $(\downarrow \text{PRED}) = \text{‘PRO’}$   
                                       $(\downarrow_{\sigma} \text{DF}) = \text{TOPIC}$   
                                       $(\downarrow \text{INDEX PERS}) = 3$   
                                       $(\downarrow \text{INDEX NUM}) = N$  — where  $N \in \{\text{SG, DU, PL}\}$

Finally, at the fourth stage the topicality constraint was reanalyzed as a definiteness constraint, see (65).

- (65)  $V_{aff}$                        $(\uparrow \text{OBJ}) = \downarrow$   
                                       $(\downarrow \text{PRED}) = \text{‘PRO’}$   
                                       $(\downarrow_{\sigma} \text{DF}) = \text{TOPIC } (\uparrow \text{OBJ DEF}) =_c +$   
                                       $(\downarrow \text{INDEX PERS}) = 3$   
                                       $(\downarrow \text{INDEX NUM}) = N$  — where  $N \in \{\text{SG, DU, PL}\}$

Dalrymple & Nikolaeva (2011) investigate differential object marking (DOM) by exploring syntactic, semantic and informational structural differences between marked and unmarked objects in a wide range of genetically and typologically different languages. As regards Uralic, they concentrate on Tundra Nenets in the Samoyedic subfamily and on Ostyak (Khanty), Vogul (Mansi) and Hungarian in the Finno-Ugric subfamily.<sup>35</sup>

Dalrymple & Nikolaeva (2011) develop a formal theory of information structure and its place in the architecture of LFG. In this theory information structure is closely related to semantic structure. It is a favourable aspect of this approach that it makes possible a simple specification of the informational structural status of an argument by providing a DF feature value in its semantic structure.

In Tundra Nenets there is only a single object function: OBJ. First and second person (pronominal) objects do not agree with the verb, just like in Hungarian, see (63) above. Third person objects optionally agree with the verb. If there is agreement, the object has the TOPIC DF, while no such function is associated with it in the absence of agreement. Dalrymple & Nikolaeva (2011) model this in the following way.

(66) Agreement with third person topical objects:

$$\begin{aligned}(\uparrow \text{OBJ PERS}) &= 3 \\ ((\uparrow \text{OBJ})_{\sigma} \text{DF}) &= \text{TOPIC}\end{aligned}$$

This specification encodes that the semantic structure contributed by the third person object is associated with the topic role in information structure.

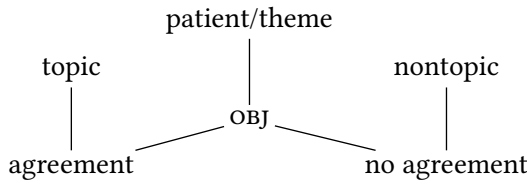
Dalrymple & Nikolaeva (2011) also distinguish a language type in which there are two object functions: OBJ and OBJ<sub>θ</sub>. They claim that Ostyak belongs to this type, in addition to Mongolian, Chatino and Hindi, among others. The OBJ<sub>θ</sub> function in these languages is only available to patient/theme arguments. Dalrymple & Nikolaeva (2011) make the following empirical generalizations. Although Ostyak has two object functions, they cannot co-occur in a sentence, because this language does not have a double object construction. In the case of verbs such as ‘give’ there are the following two possibilities: either the goal or the theme must have an OBL function. When the goal has a dative oblique function, the theme has two object choices. If it is topical, it has the agreeing OBJ function, and if it is not topical, it has the non-agreeing OBJ<sub>THEME</sub> function.

Dalrymple & Nikolaeva (2011) compare the Nenets type and the Ostyak type of DOM in the following way.

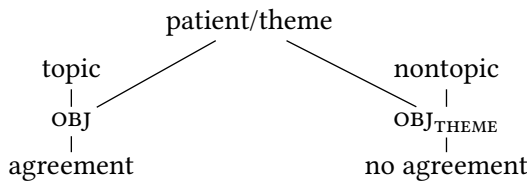
<sup>35</sup>On aspectual DOM in Estonian, see the discussion of Tamm (2006) in Section 6.2.

*Tibor Laczkó*

(67) Nenets monotransitives with third person objects:



(68) Ostyak monotransitives with patient/theme objects:



Dalrymple & Nikolaeva (2011) also point out that in the Ob-Ugric branch of Finno-Ugric languages Vogul follows the same DOM pattern as Ostyak: object marking is information structure driven: topicalization by means of object agreement. The authors hypothesize that this also held for Proto-Ob-Ugric. There are no attested semantic restrictions on agreeing objects in Ob-Ugric. As shown above, object agreement works differently in Hungarian. First and second person pronouns never trigger agreement, just like in Tundra Nenets, see above. Third person object agreement is not regulated by information structure: it is triggered by definiteness. It is only definite third person objects that trigger agreement irrespective of their discourse function status.

The authors suggest that earlier Hungarian was closer to Ostyak and Vogul, and in modern Hungarian definiteness marking is an innovation, after the development of the grammatical category of definiteness and the appearance of grammatical articles. Their reconstruction of the relevant linguistic historical processes is as follows. They assume that the Ob-Ugric system of DOM, which is exclusively based on information structure, is the most archaic type, and probably it can be hypothesized for Proto-Eastern-Uralic, i.e. the Proto-Uralic dialects from which the Samoyedic and Ugric languages developed. At a later stage, agreement became reduced to third person topical objects in Samoyedic and Proto-Hungarian as a consequence of the fact that third person was frequently associated with secondary topicality. By contrast, first and second person pronouns occupy the highest position on a scale of topic-worthiness. Dalrymple & Nikolaeva (2011) suggest that the Samoyedic languages (Nenets, Selkup and Nganasan) and Old Hungarian grammaticalized the tendency that first and second person

## 6 LFG and Finno-Ugric languages

pronouns are likely primary topics and unlikely secondary topics. Thus, they cannot correspond to the primary object, given that in these languages it tends to be strongly associated with secondary topic. No such restrictions hold for third person objects. Hungarian and (possibly) Selkup represent the next historical stage, at which the grammatical marking of third person topical objects is extended to non-topical definite objects. According to Dalrymple & Nikolaeva (2011) this change manifests the spreading of grammatical marking to non-topical objects that exhibit topic-worthy features with the concomitant loss of relatedness to information structure.<sup>36</sup>

This section on DOM has shown how complex these phenomena are in Uralic languages in general and in Finno-Ugric languages in particular. It has also demonstrated that LFG's well-developed modular architecture provides the necessary and appropriate formal devices to capture both the synchronic differences between languages and the diachronic processes in a principled manner.

## 7.2 Evidentiality

Asudeh & Toivonen (2017) propose a modular LFG approach to evidentiality, which is a well-established morpho-syntactic category in a considerable number of languages, for instance, Tariana, Cherokee, Cheyenne, Quechua and Tuyuca. These languages employ fully grammaticalized evidentiality morphology, which encodes the source and reliability of speakers' knowledge. Other languages, e.g. English, do not have such evidentiality marking, and they use alternative means to express sources of evidence or degrees of certainty about evidence (*apparently*, *I saw that...*, etc.). For the description of grammaticalized evidentiality they use the following f-structure features: [DIRECT ±], [VISUAL ±], [REPORTED ±], which also express semantic content to be captured as modifiers on events in Glue Semantics. In languages like English (with non-grammaticalized evidentiality) predicates like *sound* and *seem* optionally encode evidentiality information for the semantic component of the theory. The authors argue that LFG's modular architecture is especially well-suited to capturing the systematic similarities and

<sup>36</sup>Dalrymple & Nikolaeva make the following footnote comment. 'An alternative explanation was recently suggested by Coppock & Wechsler (2010), who claim that object agreement in proto-Uralic was initially restricted to third person topical objects. It later spread to all topical objects in Northern Ostyak and Vogul, whereas Samoyedic languages preserve the original situation. This suggestion provides an elegant analysis of feature loss as a mechanism of historical change: Northern Ostyak lost the specification that restricted topical agreement to third person objects (the (↑ OBJ PERS)=3 specification for agreeing verbs). However, the causal mechanism of this development remains unclear: it presupposes the spread of marking to unlikely contexts' (Dalrymple & Nikolaeva 2011: 201).

*Tibor Laczkó*

differences between grammaticalized and non-grammaticalized ways of expressing evidentiality across languages.

Szabó (2021) points out that in the family of Uralic languages both evidentiality systems can be found. For instance, the Finnic, the Saamic and the Mordvinian languages and Hungarian do not have grammaticalized evidentiality. By contrast, Estonian, Livonian, Mari, Komi, Udmurt as well as the entire Ob-Ugric and Samoyedic branches employ grammaticalized evidentiality.

Szabó (2021) sketches an LFG approach to grammaticalized evidentiality in Udmurt. She shows that there are two past tense paradigms in this language, and the 2<sup>nd</sup> past is used to express the source of information, among other aspects of morpho-syntax. Therefore, this verb form is multiply ambiguous. Szabó (2021: 82) captures this by proposing that the 2<sup>nd</sup> past contributes the following attribute-value pair to the f-structure of a sentence.<sup>37</sup>

(69) [SOURCE RES ∨ PFV ∨ HEAR ∨ FOLK ∨ MIR ∨ INFER ∨ NON-V]

As (69) shows, in this domain the f-structure is multiply ambiguous with all these disjunctive values for SOURCE, and the assumption is that it is basically the context that disambiguates.

Tamm (2008) shows that in Estonian partitive case-marking has either epistemic modality or aspectual use. In the former, it encodes incomplete evidence (cf. grammaticalized evidentiality marking), and in the latter, it presents an event as incomplete. The lack of partitive-marking indicates complete evidence and complete event, respectively. In this language both verbs and object arguments can be marked for partitive. Tamm proposes the lexical form in (70) for the aspectual partitive case marker on the object, and the lexical forms in (71) and (72) for the impersonal and personal evidentiality markers on verbs, respectively.

(70) (↑ CASE) = PARTITIVE  
((OBJ ↑) EVENT) ≠ COMPLETE

(71) [-ta-vat] (↑ FORM) = PARTITIVE EVIDENTIAL  
(↑ MODE OF COMMUNICATION) = INDIRECT  
(↑ EVIDENCE) ≠ COMPLETE  
(↑ VOICE) = IMPERSONAL

<sup>37</sup>Where RES = resultative, PFV = perfective, HEAR = hearsay, FOLK = folklore, MIR = mirative, INFER = inferential, NON-V = non-volitional.

- (72) [-va-t]      (↑ FORM) = PARTITIVE EVIDENTIAL  
                      (↑ MODE OF COMMUNICATION) = INDIRECT  
                      (↑ EVIDENCE) ≠ COMPLETE  
                      (↑ VOICE) = PERSONAL

Tamm sketches a Discourse Representation Theory-based semantic description associated with the f-structure representation.

For further discussions and analyses of evidentiality, see Szabó (2017) on Udmurt, and Tamm (2004c, 2012a) on Estonian. On partitives, also see Tamm (2012b).

## 8 Noun phrase phenomena in Hungarian

### 8.1 C-structure issues

As we show below, Hungarian noun phrases have been analyzed as either NPs or DPs in LFG approaches. Both views are fully legitimate in this framework, given that the standard LFG inventory of functional categories contains D (in addition to I and C).<sup>38</sup> It is a crucial property of possessive noun phrases in this language that the possessor can be expressed in either nominative or dative case, and the two variants occupy distinct syntactic positions. Despite this fact, only one of them can occur in any single possessive noun phrase, that is they are in complementary distribution, as opposed to the possible co-occurrence of *'s* and *of* possessors in English.

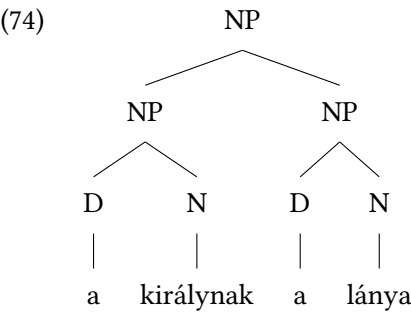
Chisarik & Payne (2003: 189) use an NP approach to the representation of Hungarian and English noun phrases, see the structures they assume for (73) and (75)<sup>39</sup> in (74) and (76), respectively.

- (73) Hungarian:  
       a király-nak a lány-a  
       the king-DAT the daughter-POSS.3SG  
       ‘the king’s daughter’

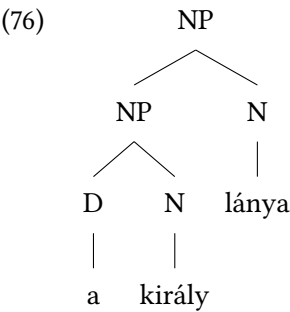
<sup>38</sup>It is not unusual to find alternative categorial analyses of the same construction types in LFG. For instance Bresnan (2001) treats finite English sentences that contain no auxiliaries (e.g. *Mary opened the door*) as having the category S, while Dalrymple (2001) employs an IP approach.

<sup>39</sup>Notice that Hungarian possessive noun phrases belong to the head-marking type.

*Tibor Laczkó*



- (75) Hungarian:  
a király lány-a  
the king.NOM daughter-POSS.3SG  
'the king's daughter'



They provide the following justifications for these representations. On the one hand, the dative possessor, see (74), can function as a predeterminer to coordinated NPs as in their example in (77).

- (77) Hungarian:  
a király-nak [NP [ a fi-a ] és [ a lány-a ]]  
the king-DAT the son-POSS.3SG and the daughter-POSS.3SG  
'the king's son and daughter'

On the other hand, the nominative possessor stands in complementary distribution with the definite article, just like the 's possessor in English.

The following remarks can be made on this approach. First, the coordination facts can also be captured in a DP analysis in which the dative possessor is in Spec,DP and Chisarik and Payne's NP is a D', where the definite article is the

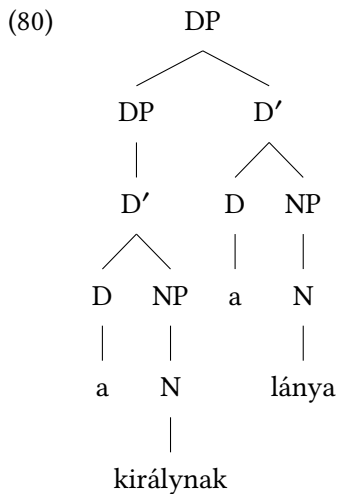


D head and the other constituent is (the head of) an NP.<sup>40</sup> Second, it would need some justification to assume that a word-level functional category (D) is in complementary distribution with a phrasal category (NP).<sup>41</sup> Third, in the case of pronominal nominative possessors there is no complementary distribution with the definite article; moreover, they must co-occur, compare (78) and (79).<sup>42</sup>

- (78) Hungarian:  
 (\*a) János lány-a  
 the John.NOM daughter-POSS.3SG  
 'John's daughter'

- (79) Hungarian:  
 \*(az) ő lány-a  
 the he.NOM daughter-POSS.3SG  
 'his daughter'

Motivated by Szabolcsi's (1994) seminal GB analysis, Laczkó in Laczkó (1995) and all subsequent work adapts a DP approach.<sup>43</sup> The essential aspects of his structural representation of (73) and (75) would be as in (80) and (81), respectively.



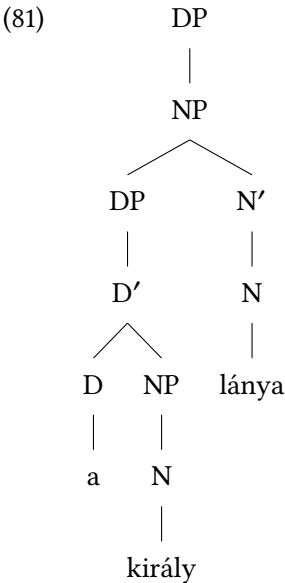
<sup>40</sup>See Laczkó's (1995) DP structure in (80) below.

<sup>41</sup>It seems to be a further minor complication that the functional category D is used in an unusual way: it does not head and project a DP.

<sup>42</sup>(78) shows the grammaticality properties of this construction type in standard Hungarian. However, Szabolcsi (1994) documents a dialectal version in which even such non-pronominal nominative possessor constructions follow the pattern exemplified in (79).

<sup>43</sup>Without adopting theory-specific details like moving the nominative possessor from Spec,NP to Spec,DP, where it acquires dative case, as in Szabolcsi's (1994) GB analysis.

*Tibor Laczkó*



This approach avoids the complications mentioned in connection with Chisarik & Payne’s (2003) NP analysis.

## 8.2 Event nominalization

### 8.2.1 Argument structure inheritance

Following Grimshaw (1990) and Szabolcsi (1994), among others, Laczkó in Laczkó (1995) and in all relevant subsequent work assumes that complex event nominals (CENS) derived by the *-ás/-és* suffix (henceforth: *Ás* suffix) inherit the argument structure of the input verb, as opposed to simple event nouns (SENS) and result nouns (RESES). The most important properties of Hungarian CENS are as follows; see also Laczkó (2000b, 2003, 2009a).

When an *Ás* noun has both a simplex form and a complex form containing a perfectivizing preverb, the latter is always a CEN and the former is very often ambiguous: CEN vs. SEN and/or RES. Compare the examples in (82).

(82) Hungarian:

- a. Anna        vizsgáztat-ás-a  
 Anne.NOM examine-ÁS-POSS.3SG  
 ‘Anne’s examination’  
 CEN: Anne = patient  
 SEN: Anne = examiner or examinee

- b. Anna le-vizsgáztat-ás-a (a professzor által)  
 Anne.NOM PFV-examine-ÁS-POSS.3SG the professor by  
 ‘the examination of Anne (by the professor)’  
 CEN: Anne = patient
- c. Anna vizsgá-ja  
 Anne.NOM exam-POSS.3SG  
 ‘Anne’s exam’  
 SEN: Anne = examiner or examinee

(82a) contains a derived nominal without a perfectivizing preverb, and it can be used as either a CEN with an argument structure or as a SEN without an argument structure (with only a lexical conceptual structure). In the former use *Anna* is interpreted as the patient argument of the nominal predicate, in the latter use it is interpreted as a participant in an examination situation, whether the examiner or the examinee. By contrast, in (82b) the derived nominal contains a perfectivizing preverb, and it can only be analysed as a CEN with obligatory argument structure and *Anna* must be interpreted as the patient argument. In (82c) the head is an underived noun and it can only be a SEN.

The expression of the arguments of the derived nominal predicate is as obligatory as in the case of the input verb.

(83) Hungarian:

A vizsgáztat-ás két órá-ig tart-ott.  
 the examine-ÁS-POSS.3SG two hour-for last-PST.3SG  
 ‘The examination lasted for two hours.’ (SEN)

(84) Hungarian:

\*A le-vizsgáztat-ás két órá-ig tart-ott.  
 the PFV-examine-ÁS-POSS.3SG two hour-for last-PST.3SG  
 ‘The examination lasted for two hours.’ (CEN)

As (83) shows, when no complement is present, an otherwise ambiguous (CEN/SEN) nominal must be interpreted as a SEN. (84) demonstrates that an ‘only CEN’ nominal cannot occur without its obligatory internal argument(s). The external argument can be suppressed optionally, see (82b) above.

CENS cannot be pluralized, see (85).

(85) Hungarian:

\*Anna le-vizsgáztat-ás-a-i  
 Anne.NOM PFV-examine-ÁS-POSS.3SG-PL  
 ‘\*the examinations of Anne’ (CEN)

Tibor Laczkó

When an adjunct in the DP with a derived nominal head is expressed by a postpositional phrase, this PP has to be ‘adjectivalized’ either by combining it with a formative element, one of the present participial forms of the copula: *való* ‘being’, glossed as VALÓ, or by attaching the adjectivizing suffix *-i* (glossed as AFF) to the postposition. In such cases, the VALÓ version is only compatible with the CEN reading of an otherwise ambiguous nominal predicate, while the *-i* variant retains the ambiguity, cf. (86a) and (86b). This is Szabolcsi’s (1994) famous *való*-test for unambiguously identifying CENS in Hungarian.<sup>44</sup>

(86) Hungarian:

- a. az ebéd után való beszélget-és  
the lunch after VALÓ converse-ÁS  
‘conversing after lunch’ (CEN)
- b. az ebéd után-i beszélget-és  
the lunch after-AFF converse-ÁS  
‘conversing after lunch’ (CEN)  
‘the conversation after lunch’ (SEN)

The core arguments of CENS can receive a variety of [–r] GFS in several LFG approaches to Hungarian, see Section 8.2.2. Non-core arguments are typically expressed by case-marked DPs and postpositional phrases, and they are mapped onto OBL functions. Adjuncts can also be expressed by case-marked DPs and PPs. In addition, they can be realized by APs, especially when the input verb would take an AdvP for the same kind of modification, e.g. *váratlan-ul* ‘unexpectedly’ (Adv) vs. *váratlan* ‘unexpected’ (A). For empirical generalizations about the major (structural and categorial) ways of realizing OBL and ADJUNCT functions in CEN constructions and LFG analyses, see Laczkó (1995, 2003).

The Hungarian event nominalization phenomena presented above are relevant for theorizing in generative grammar in general and in LFG in particular for the following reasons. Grimshaw’s (1990) influential proposal substantially distinguishing CENS from SENS and RESES is based on English data, primarily on *-tion* nominalization. In this language, however, these derived nouns are genuinely ambiguous and, therefore, it is often difficult to employ Grimshaw’s diagnostics, e.g. (non-)pluralizability, to definitely tell the CEN and SEN uses apart. Due to this fact, Grimshaw’s theory has been criticized from a variety theoretical perspectives, see Laczkó (2000b) and the references therein. By contrast, in Hungarian there are clear morphological and syntactic indicators, and the diagnostics can

<sup>44</sup> Also see Laczkó & Rákosi (2007).

be applied reliably and unambiguously. This situation has motivated some LFG practitioners to investigate event nominalization thoroughly and, among other things, to develop various LMT analyses of argument realization in this domain, see Section 8.2.2.

8.2.2 Functional issues

A variety of inventories of GFS in Hungarian DPs with CEN heads and a consequential variety of LMT analyses have been proposed, see Table 6.<sup>45</sup>

Table 6: GFS in Hungarian DPs

	Laczko (1995)	Chisarik & Payne (2003)	Laczko (2004)
DP <sub>DAT</sub>	POSS	SUBJ	SUBJ/POSS
DP <sub>NOM</sub>	POSS	NCOMP	SUBJ/POSS
DP <sub>OBL</sub> /PP	OBL	OBL	OBL

Laczko (1995) uses GFS standardly employed in noun phrases (POSS and OBL). Assuming that POSS is a semantically unrestricted function, he develops an LMT approach in which there is a POSS Condition that is the nominal domain counterpart of the SUBJ Condition in the verbal domain. The SUBJ Condition requires that every (verbal) predicator must have a Subject, see Bresnan (1990), for instance. Laczko’s (1995: 85) POSS Condition states: ‘every event nominal predicator must have a Possessor’.

Rather exceptionally in the generative literature on Hungarian noun phrases, Chisarik & Payne (2003) assume that the two possessor constituents bear distinct GFS, both of which are taken to be semantically unrestricted. The dative realizes the SUBJ function in the nominal domain, while the nominative expresses a new, DP-specific function: NCOMP. SUBJ is considered to be discourse-related, while NCOMP is not.

Laczko (2004) assumes that both the dative possessor and the nominative possessor can overtly realize either the SUBJ or the POSS GFS, both of which are regarded as semantically unrestricted. Furthermore, the SUBJ argument can also be expressed by an LFG-style PRO. Given this nature and distribution of these GFS, Laczko’s LMT analysis can adopt the SUBJ Condition from the verbal domain. In addition, his approach can formally handle (anaphoric) control into possessive DPs in Hungarian with the standard LFG mechanism even in the case of CENS

<sup>45</sup>Charters (2014) proposes a new DF in Hungarian possessive DPs: ANCHOR.

*Tibor Laczkó*

derived from transitive verbs, which Laczkó's (1995) system cannot do. Consider the following examples.

(87) Hungarian:

- a. Péter elkezdte a kiabál-ás-t.  
Peter.NOM started the shout-ÁS-ACC  
'Peter started the shouting.'
- b. Péter elkezdte a dal énekl-és-é-t.  
Peter.NOM started the song.NOM sing-ÁS-POSS.3SG-ACC  
'Peter started the singing of the song.'

In Laczkó's (1995) system, the f-structure of the DP in (87a) contains a POSS PRO, which is anaphorically controlled by the matrix subject, and in (87b) *a dal* 'the song' has the POSS function, and (in the absence of any other available GF for the agent controllee) Laczkó is forced to assume that control takes place in a different dimension. By contrast, in Laczkó's (2004) approach there is a controlled PRO SUBJ in both cases, and in (87b) *a dal* 'the song' has the POSS function. Laczkó's (2004) SUBJ & POSS theory receives further independent support from Laczkó & Rákosi (2019), who argue that this GF inventory is necessary for the adequate LFG handling of certain binding facts in Hungarian DPs. Laczkó (2008b, 2009b), in response to Kenesei (2005), proposes that both T participial constructions and CEN constructions should have a dual PRO & suppression analysis for an adequate treatment of binding and control phenomena.

### 8.3 Possessives

#### 8.3.1 Finnish

Toivonen (2000) develops an analysis of the morpho-syntax of Finnish possessive noun phrases. This language has the widely attested POSS pro-drop in the case of first and second person possessors, see a 1SG example in (88), and Toivonen's lexical representation of the pronoun and the possessive suffix (glossed as POSS) in (89) and (90), respectively.

- (88) Finnish:  
Pekka näkee (minun) ystävä-ni.  
Pekka sees my friend-POSS.1SG  
'Pekka sees my friend.'

$$(89) \quad \text{minun:} \quad \left[ \begin{array}{c} \text{POSS} \left[ \begin{array}{c} \text{PRED 'PRO'} \\ \text{PERS 1} \\ \text{NUM SG} \end{array} \right] \end{array} \right]$$

$$(90) \quad \text{-ni:} \quad \left[ \begin{array}{c} \text{POSS} \left[ \begin{array}{c} (\text{PRED 'PRO'}) \\ \text{PERS 1} \\ \text{NUM SG} \end{array} \right] \end{array} \right]$$

In the third person there is an interesting split between the possessive pronoun and the possessive suffix when the latter provides the PRED feature (i.e. in the case of pro-drop). The pronoun must not be bound by the matrix subject, while the POSS-PRO must, cf. (91) and (92).

- (91) Finnish:  
 Pekka näkee hänen ystävä-nsä.  
 Pekka sees his/her friend-POSS.3SG  
 'Pekka sees his/her<sub>i/j</sub> friend.'

- (92) Finnish:  
 Pekka näkee ystävä-nsä.  
 Pekka sees friend-POSS.3SG  
 'Pekka sees his/her<sub>i/\*j</sub> friend.'

Furthermore, the 3SG.POSS suffix cannot agree with a non-human possessor:

- (93) Finnish:  
 sen ruokaa(\*-nsa)  
 its food-POSS.3SG  
 'its food'

Toivonen captures these facts by means of the following lexical forms.<sup>46</sup>

$$(94) \quad \text{hänen:} \quad \left[ \begin{array}{c} \text{POSS} \left[ \begin{array}{c} \text{PRED 'PRO'} \\ \text{PERS 3} \\ \text{GEND HUM} \\ \text{NUM SG} \\ \text{SB —} \end{array} \right] \end{array} \right]$$

$$(95) \quad \text{pron. -nsA:} \quad \left[ \begin{array}{c} \text{POSS} \left[ \begin{array}{c} \text{PRED 'PRO'} \\ \text{PERS 3} \\ \text{SB +} \end{array} \right] \end{array} \right]$$

<sup>46</sup>SB stands for obligatorily subject bound.

*Tibor Laczkó*

- (96) *agr. -nsA:*       $\left[ \text{POSS} \left[ \text{PERS } 3 \right] \right], \text{GEND}=\text{c HUM}$

Toivonen also compares corresponding possessive noun phrase constructions in Estonian and Northern Saami. Toivonen (2001) provides a historical context for her analysis in Toivonen (2000), and she also discusses dialectal variation in Finnish with respect to these phenomena. Her proposal involves the erosion of features other than PRED ‘pro’, which makes it very similar to Coppock & Wechsler’s (2010) analysis of Ostyak and Hungarian in Section 7.1.2.

### 8.3.2 Hungarian

Laczkó (2001) develops an LFG approach to the inflectional phenomena in Hungarian possessive DPs in the spirit of Item and Arrangement morphology.<sup>47</sup> Consider the following examples.

- (97) Hungarian:
- a. a toll-a-i-nk  
the pen-POSS-PL-1PL  
‘our pens’
  - b. a toll-a-i  
the pen-POSS-PL.3SG  
‘her pens’
  - c. a toll-a  
the pen-POSS.3SG  
‘her pen’
  - d. a hajó-i  
the ship-POSS.PL.3SG  
‘her ships’

Laczkó postulates the following sets of functional annotations in the lexical forms of *-a* and *-i*, the main point being that the same morphological form (morph) can encode fewer or more features depending on what other morphs it is combined with, see the optional features in (98).

- (98) a. *-a*       $(\uparrow \text{POSS})$       [97a, 97b]  
                   $(\uparrow \text{POSS PERS}) = 3$       [97]  
                   $(\uparrow \text{POSS NUM}) = \text{SG}$   
                   $((\uparrow \text{POSS PRED}) = \text{‘PRO’})$

<sup>47</sup>By contrast, Laczkó (2018) proposes a Word and Paradigm approach, arguing that it has considerable implementational advantages.



- b. *-i*      (↑ POSS)      [97a]  
               (↑ NUM)  
               (↑ POSS PERS) = 3      [97b, 97d]  
               (↑ POSS NUM) = SG  
               ((↑ POSS PRED) = ‘PRO’)

## 9 Further reading

Limitations of space have prevented us from discussing additional phenomena in Finno-Ugric languages and their analyses. Below we provide references to further works that we recommend to the interested reader.

On predicate-argument relationships in Hungarian, see Komlósy (1992, 1994) and Rákosi (2008). On causatives in Hungarian, see Komlósy (2000). On argument realization alternations in Finnish and Estonian, see Ackerman & Moore (1999), in Hungarian, see Ackerman (1992) and Laczkó (2013b). On Uralic conjugation classes and verbs imposing restrictions on argument structure, see Abondolo (1998), Nikolaeva (2014) and Tamm & Vainikka (to appear). On argument vs. (thematic) adjunct issues in Hungarian, see Rákosi (2003, 2006a,b, 2012). On the grammaticalization of the Estonian perfective particles, see Tamm (2004b). On scalar verb classes, aspect and partitive and total case assignment in Estonian, see Tamm (2012c). On the pragmatics of morphological case in the verbal domain of Finnic languages, see Tamm (2011b). On Estonian object and adverbial case marking with verbs of motion, see Tamm (2007b). On case and aspectuality in Estonian, see Tamm (2008, 2012b). On raising and equi constructions in Estonian, see Tamm (2004c, 2008). On a variety of analyses of *wh*-questions in Hungarian, see Mycock (2004, 2006, 2010, 2013), Gazdik (2010) and Laczkó (2014b, to appear). On two Finno-Ugric contributions to the COMP debate in LFG,<sup>48</sup> see Belyaev et al. (2017) for COMP on the basis of Moksha Mordvin phenomena and Szűcs (2018a) against COMP on the basis of Hungarian facts. On ‘operator raising’ in Hungarian, see Coppock (2003) and Szűcs (2013, 2014, 2018b). On binding and control relations of anaphors in Hungarian, see Rákosi (2009, 2010), Laczkó & Rákosi (2019), Szűcs (2019) and Laczkó et al. (2020). On reflexivity and binding in Uralic languages, see Volkova (2014, to appear). On participial constructions in Hungarian, see Komlósy (1992, 1994) and Laczkó (1995, 2000a, 2005). On derived and inherent relational nouns in Hungarian, see Laczkó (2008a, 2009b). On elliptical noun phrases in Hungarian, see Laczkó (2007). On modelling (in)definiteness

<sup>48</sup>For instance, see Dalrymple & Lødrup (2000) and Lødrup (2012) for COMP, and Alsina et al. (2005) and Patejuk & Przepiórkowski (2016) against COMP, and the references in these papers.

*Tibor Laczkó*

and (typological) variation in Hungarian possessive DPs, see Laczkó (2017). On a special system of person and number marking in possessive noun phrases in Northern Ostyak, see Ackerman & Nikolaeva (1997). On natural and accidental coordination in Finnish noun phrases, see King & Dalrymple (2004) and Dalrymple & Nikolaeva (2006). On a lexical analysis of a Hungarian phrasal adjectival derivational suffix, see Laczkó (1997). On extraction from partitive DPs in Hungarian, see Chisarik (2002).

## 10 Conclusion

In this chapter we have discussed some salient, sometimes competing, LFG analyses of a variety of (morpho-)syntactic phenomena in Finno-Ugric languages, with occasional glimpses at alternative generative approaches, on the one hand, and at some related phenomena in languages belonging to Samoyedic, the other major branch of Uralic languages, on the other hand. We have dealt with clausal c-structure representational issues, verbal modifiers, focused constituents, negation, copula constructions, argument realization, subject-verb agreement, differential object marking, evidentiality and a set of noun phrase phenomena related to event nominalization.

On the basis of the interim conclusions at the end of various sections, we can make the following overall concluding remarks at the end of this chapter. On the one hand, LFG provides an appropriate and suitably flexible formal apparatus for a principled analysis of all the phenomena in all the Finno-Ugric languages discussed here. The range of these phenomena is considerably wide and varied, see above, containing several cases that pose serious challenges for generative grammar at large, for instance, the treatment of complex predicates, negation, copula constructions, discourse functions, agreement and event nominalization. On the other hand, the analysis of some of these phenomena can also contribute to LFG-internal theorizing, see, for instance, the choice between LFG treatments of complex predicates involving pvc's and clause negation.

## Acknowledgements

We are grateful to our three anonymous reviewers, who at a later stage identified themselves as Anne Tamm, Ida Toivonen and Péter Szűcs. The comments of all our reviewers have greatly contributed to improving both the content and the presentational aspects of this chapter. Our special thanks go to Anne Tamm for generously providing us with a large amount of information about phenomena

in several Uralic languages that are comparable to the phenomena whose LFG analyses are discussed in this chapter and for calling our attention to a great number of additional relevant works on Uralic languages. As usual, all remaining errors are our sole responsibility.

## Abbreviations

Besides the abbreviations from the Leipzig Glossing Conventions, this chapter uses the following abbreviations.

ADE	adessive case (marker)	PAR	partitive case
ÄRA	Estonian particle	TOT	total case
ÁS	Hungarian event nominalizer suffix	VALÓ	Hungarian adjectivalizing participle
INE	inessive case	VM	verbal modifier

## References

- Abondolo, Daniel (ed.). 1998. *The Uralic languages* (Routledge Language Family Descriptions). London/New York: Routledge.
- Ackerman, Farrell. 1987. *Miscreant morphemes: Phrasal predicates in Ugric*. Berkeley: University of California-Berkeley. (Doctoral dissertation).
- Ackerman, Farrell. 1990. The morphological blocking principle and oblique pronominal incorporation in Hungarian. In Katarzyna Dziwirek, Patrick M. Farrell & Errapel Mejías-Bikandi (eds.), *Grammatical relations: A cross-theoretical perspective*, 1–19. Stanford, CA: CSLI Publications.
- Ackerman, Farrell. 1992. Complex predicates and morpholexical relatedness: Locative alternation in Hungarian. In Ivan A. Sag & Anna Szabolcsi (eds.), *Lexical matters* (CSLI Lecture Notes), 55–83. Stanford, CA: CSLI Publications.
- Ackerman, Farrell. 2003. Lexeme derivation and multiword predicates in Hungarian. *Acta Linguistica Hungarica* 50. 7–32. DOI: 10.1556/aling.50.2003.1-2.2.
- Ackerman, Farrell & Philip Lesourd. 1997. Toward a lexical representation of phrasal predicates. In Alex Alsina, Joan Bresnan & Peter Sells (eds.), *Complex predicates*, 67–106. Stanford, CA: CSLI Publications.
- Ackerman, Farrell & John Moore. 1999. Telic entity as a proto-property of lexical predicates. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '99 conference*. Stanford, CA: CSLI Publications.

Tibor Laczkó

- Ackerman, Farrell & Irina Nikolaeva. 1997. Identity in form, difference in function: The person/number paradigm in W. Armenian and N. Ostyak. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '97 conference*. Stanford, CA: CSLI Publications.
- Ackerman, Farrell, Gregory T. Stump & Gert Webelhuth. 2011. Lexicalism, periphrasis, and implicative morphology. In Robert D. Borsley & Kersti Börjars (eds.), *Non-transformational syntax: Formal and explicit models of grammar*, 325–358. Oxford: Wiley-Blackwell. DOI: 10.1002/9781444395037.ch9.
- Alsina, Alex. 1992. On the argument structure of causatives. *Linguistic Inquiry* 23. 517–555.
- Alsina, Alex. 1997. Causatives in Bantu and Romance. In Alex Alsina, Joan Bresnan & Peter Sells (eds.), *Complex predicates*, 203–246. Stanford, CA: CSLI Publications.
- Alsina, Alex, Joan Bresnan & Peter Sells (eds.). 1997. *Complex predicates*. Stanford, CA: CSLI Publications.
- Alsina, Alex, K. P. Mohanan & Tara Mohanan. 2005. How to get rid of the COMP. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '05 conference*, 21–41. Stanford, CA: CSLI Publications.
- Andrews, Avery D. 1990. Unification and morphological blocking. *Natural Language & Linguistic Theory* 8(4). 507–557. DOI: 10.1007/bf00133692.
- Asudeh, Ash & Ida Toivonen. 2017. A modular approach to evidentiality. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '17 conference*, 45–65. Stanford, CA: CSLI Publications.
- Asztalos, Erika. 2020. Focus in Udmurt: Positions, contrastivity, and exhaustivity. *Finno-Ugric Languages and Linguistics* 9. 14–57.
- Attia, Mohammed. 2008. A unified analysis of copula constructions in LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 89–108. Stanford, CA: CSLI Publications.
- Belyaev, Oleg, Anastasia Kozhemyakina & Natalia Serdobolskaya. 2017. In defense of COMP: Complementation in Moksha Mordvin. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '17 conference*, 83–103. Stanford, CA: CSLI Publications.
- Börjars, Kersti, Erika Chisarik & John Payne. 1999. On the justification for functional categories in LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '99 conference*. Stanford, CA: CSLI Publications.
- Brattico, Pauli. 2019. Word order in Finnish: Nonconfigurationality, movement or adjunction? *Finno-Ugric Languages and Linguistics* 8(2). 2–26.

- Bresnan, Joan. 1982. The passive in lexical theory. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 3–86. Cambridge, MA: The MIT Press.
- Bresnan, Joan. 1990. Monotonicity and the theory of relation-changes in LFG. *Language Research* 26(4). 637–652.
- Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.
- Bresnan, Joan & Sam A. Mchombo. 1987. Topic, pronoun, and agreement in Chicheŵa. *Language* 63. 741–782. DOI: 10.2307/415717.
- Brody, Michael. 1990. Remarks on the order of elements in the Hungarian focus field. In István Kenesei (ed.), *Approaches to Hungarian, volume 3: Structures and arguments*, 95–122. Szeged: JATE.
- Butt, Miriam. 2003. The light verb jungle. In C. Quinn, C. Bower & G. Aygen (eds.), *Papers from the Harvard/Dudley House light verb workshop* (Harvard Working Papers in Linguistics 9), 1–49.
- Butt, Miriam & Tracy Holloway King. 2004. The status of case. In Veneeta Dayal & Anoop Mahajan (eds.), *Clause structure in South Asian languages*. Berlin: Springer Verlag. DOI: 10.1007/978-1-4020-2719-2.
- Butt, Miriam, Tracy Holloway King, María-Eugenia Niño & Frédérique Segond. 1999. *A grammar writer's cookbook*. Stanford, CA: CSLI Publications.
- Charters, Helen. 2014. Anchor: A DF in DP. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 200–220. Stanford, CA: CSLI Publications.
- Chisarik, Erika. 2002. Partitive noun phrases in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '02 conference*, 96–115. Stanford, CA: CSLI Publications.
- Chisarik, Erika & John Payne. 2003. Modelling possessor constructions in LFG: English and Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Nominals: Inside and out*, 181–199. Stanford, CA: CSLI Publications.
- Coppock, Elizabeth. 2003. Sometimes it's hard to be coherent. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '03 conference*, 126–143. Stanford, CA: CSLI Publications.
- Coppock, Elizabeth & Stephen Wechsler. 2010. Less-travelled paths from pronoun to agreement: The case of the Uralic objective conjugations. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 165–185. Stanford, CA: CSLI Publications.
- Coppock, Elizabeth & Stephen Wechsler. 2012. The objective conjugation in Hungarian: Agreement without phi-features. *Natural Language and Linguistic Theory* 30(30). 699–740. DOI: 10.1007/s11049-012-9165-5.

Tibor Laczkó

- Dalrymple, Mary. 2001. *Lexical Functional Grammar*. Vol. 34 (Syntax and Semantics). New York: Academic Press. DOI: 10.1163/9781849500104.
- Dalrymple, Mary, Helge Dyvik & Tracy Holloway King. 2004. Copular complements: Closed or open? In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*, 188–198. Stanford, CA: CSLI Publications.
- Dalrymple, Mary & Helge Lødrup. 2000. The grammatical functions of complement clauses. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '00 conference*, 104–121. Stanford, CA: CSLI Publications.
- Dalrymple, Mary & Irina Nikolaeva. 2006. Syntax of natural and accidental coordination: evidence from agreement. *Language* 82(4). 824–849. DOI: 10.1353/lan.2006.0189.
- Dalrymple, Mary & Irina Nikolaeva. 2011. *Objects and information structure* (Cambridge Studies in Linguistics). Cambridge, UK: Cambridge University Press.
- Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://wals.info/>.
- Forst, Martin, Tracy Holloway King & Tibor Laczkó. 2010. Particle verbs in computational LFGs: Issues from English, German, and Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 228–248. Stanford, CA: CSLI Publications.
- Gazdik, Anna. 2010. Multiple questions in French and in Hungarian: An LFG account. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 249–269. Stanford, CA: CSLI Publications.
- Gazdik, Anna. 2012. Towards an LFG analysis of discourse functions in Hungarian. In Ferenc Kiefer & Zoltán Bánréti (eds.), *Twenty years of theoretical linguistics in Budapest*, 59–92. Hungarian Academy of Sciences, Research Institute for Linguistics & Tinta Publishing House.
- Gazdik, Anna & András Komlósy. 2011. On the syntax-discourse interface in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '11 conference*, 215–235. Stanford, CA: CSLI Publications.
- Grimshaw, Jane. 1990. *Argument structure*. Cambridge, MA: The MIT Press.
- Groot, Casper de. 2017. *Uralic essive and the expression of impermanent state*. Vol. 119 (Typological Studies in Language). Amsterdam: John Benjamins.
- Hiietam, Katrin. 2003. *Definiteness and grammatical relations in Estonian*. Manchester: University of Manchester. (Doctoral dissertation).
- Kenesei, István. 2005. Nonfinite clauses in derived nominals. In Christopher Piñón & Péter Siptár (eds.), *Approaches to Hungarian, volume 9: Papers from the Düsseldorf conference*, 159–186. Budapest: Akadémiai Kiadó.

- Kiefer, Ferenc & László Honti. 2003. Verbal “prefixation” in the Uralic languages. *Acta Linguistica Hungarica* 50. 137–153. DOI: 10.1556/aling.50.2003.1-2.8.
- King, Tracy Holloway. 1995. *Configuring topic and focus in Russian*. Stanford, CA: CSLI Publications.
- King, Tracy Holloway & Mary Dalrymple. 2004. Determiner agreement and noun conjunction. *Journal of Linguistics* 4. 69–104. DOI: 10.1017/s0022226703002330.
- É. Kiss, Katalin. 1992. Az egyszerű mondat szerkezete [The structure of the simple sentence]. In Ferenc Kiefer (ed.), *Strukturális magyar nyelvtan 1: Mondattan [Structural Hungarian grammar 1: Syntax]*, 79–177. Budapest: Akadémiai Kiadó.
- É. Kiss, Katalin. 1994. Sentence structure and word order. In Ferenc Kiefer & Katalin É. Kiss (eds.), *The syntactic structure of Hungarian*, vol. 27 (Syntax and Semantics), 1–90. Academic Press.
- É. Kiss, Katalin. 1995. *Discourse configurational languages*. Oxford: Oxford University Press.
- É. Kiss, Katalin. 2004. Egy igekötoelmélet vázlatja [Outlines of a theory of verbal particles]. *Magyar Nyelv* 50. 15–43.
- É. Kiss, Katalin. 2020. Accusative or possessive? The suffix of pronominal objects in Ob-Ugric. *Finno-Ugric Languages and Linguistics* 9(1-2). 3–13.
- Komlósy, András. 1992. Régensek és vonzatok [Predicates and complements]. In Ferenc Kiefer (ed.), *Strukturális magyar nyelvtan 1: Mondattan [Structural Hungarian grammar 1: Syntax]*, 299–527. Budapest: Akadémiai Kiadó.
- Komlósy, András. 1994. Complements and adjuncts. In Ferenc Kiefer & Katalin É. Kiss (eds.), *The syntactic structure of Hungarian*, vol. 27 (Syntax and Semantics), 91–178. Academic Press.
- Komlósy, András. 2000. A műveltetés [Causatives]. In Ferenc Kiefer (ed.), *Strukturális magyar nyelvtan 3: Morfológia [Structural Hungarian grammar 3: Morphology]*, 215–292. Budapest: Akadémiai Kiadó.
- Komlósy, András. 2001. *A lexikai-funkcionális grammatika mondattanának alapfogalmai [Basic concepts of the syntax of lexical functional grammar]*. Budapest: Tinta Kiadó.
- Laczkó, Tibor. 1989. A lexikai-funkcionális grammatika főbb jellemzői [Major traits of lexical-functional grammar]. *Általános Nyelvészeti Tanulmányok* 17. 367–374.
- Laczkó, Tibor. 1995. *The syntax of Hungarian noun phrases – A lexical-functional approach* (Metalinguistica 2). Frankfurt am Main: Peter Lang.
- Laczkó, Tibor. 1997. An analysis of -ú/-ű adjectives in Hungarian: The case of another morphologically bound predicate. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '97 conference*. Stanford, CA: CSLI Publications.



Tibor Laczkó

- Laczkó, Tibor. 2000a. A melléknévi és határozói igenévképzők [Adjectival and adverbial participles]. In Ferenc Kiefer (ed.), *Strukturális magyar nyelvtan 3: Morfológia* [*Structural Hungarian grammar 3: Morphology*], 409–452. Budapest: Akadémiai Kiadó.
- Laczkó, Tibor. 2000b. Derived nominals, possessors, and lexical mapping theory in Hungarian DPs. In Miriam Butt & Tracy Holloway King (eds.), *Argument realization*, 189–227. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2001. A magyar birtokos DP inflexiók morfológiájáról lexikai-funkcionális megközelítésben [Remarks on the inflectional morphology of Hungarian possessive DPs from a lexical-functional perspective]. In Marianne Bakró-Nagy, Zoltán Bánréti & Katalin É. Kiss (eds.), *Újabb tanulmányok a strukturális magyar nyelvtan és a nyelvtöréből. kiefer ferenc tiszteletére barátai és tanítványai* [*New studies on Hungarian structural grammar and historical linguistics. A festschrift for Ferenc Kiefer by his friends and students*], 59–77. Budapest: Osiris Kiadó.
- Laczkó, Tibor. 2003. On oblique arguments and adjuncts in Hungarian event nominals. In Miriam Butt & Tracy Holloway King (eds.), *Nominals: Inside and out*, 201–234. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2004. Grammatical functions, LMT, and control in the Hungarian DP revisited. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*, 313–333. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2005. Nominalization, participle-formation, typology and lexical mapping. In Christopher Piñón & Péter Siptár (eds.), *Approaches to Hungarian, volume 9: Papers from the Düsseldorf conference*, 205–230. Budapest: Akadémiai Kiadó.
- Laczkó, Tibor. 2007. On elliptical noun phrases in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '07 conference*, 323–342. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2008a. A relációs főnevek [Relational nouns]. In Ferenc Kiefer (ed.), *Strukturális magyar nyelvtan 4: A szótár szerkezete* [*Structural Hungarian grammar 4: The structure of the lexicon*], 323–503. Budapest: Akadémiai Kiadó.
- Laczkó, Tibor. 2008b. On binding, empty categories, and morphosyntactic processes in “passive” participial constructions. In Christopher Piñón & Szilárd Szentgyörgyi (eds.), *Papers from the Veszprém conference: Approaches to Hungarian 10*, 103–126. Budapest: Akadémiai Kiadó.
- Laczkó, Tibor. 2009a. On the -Ás suffix: Word formation in the syntax? *Acta Linguistica Hungarica* 56(1). 23–114. DOI: 10.1556/aling.56.2009.1.2.



- Laczkó, Tibor. 2009b. Relational nouns and argument structure: Evidence from Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 399–419. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2012. On the (un)bearable lightness of being an LFG style copula in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 341–361. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2013a. Hungarian particle verbs revisited: Representational, derivational and implementational issues from an LFG perspective. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '13 conference*, 377–397. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2013b. On presenting something in English and Hungarian. In Tracy Holloway King & Valeria de Paiva (eds.), *From quirky case to representing space: Papers in honor of Annie Zaenen*, 181–194. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2014a. An LFG analysis of verbal modifiers in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 346–366. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2014b. Essentials of an LFG analysis of Hungarian finite sentences. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 325–345. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2014c. Outlines of an LFG-XLE account of negation in Hungarian sentences. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 304–324. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2015. On negative particles and negative polarity in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '15 conference*, 166–186. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2017. Modelling (in)definiteness, external possessors and (typological) variation in Hungarian possessive DPs. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '17 conference*, 243–263. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2018. Modelling possession and agreement in Hungarian DPs: A paradigmatic approach. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '18 conference*, 227–247. Stanford, CA: CSLI Publications.
- Laczkó, Tibor. 2020. Egy unortodox GB-modell hatása egy ortodox LFG-modellre [The influence of an unorthodox GB model on an orthodox LFG model]. In Éva Dékány, Tamás Halm & Balázs Surányi (eds.), *Általános Nyelvészeti Tanulmányok* 32, 155–169. Budapest: Akadémiai Kiadó.

Tibor Laczkó

- Laczkó, Tibor. to appear. *Lexicalizing clausal syntax: The interaction of syntax, the lexicon and information structure in Hungarian* (Current Issues in Linguistic Theory 354). Amsterdam: John Benjamins.
- Laczkó, Tibor & György Rákosi. 2007. Mire *való*? Egy lexikai-funkcionális esettanulmány [What is *való* for? A lexical-functional case study]. In Gábor Alberti & Ágota Fóris (eds.), *A mai magyar formális nyelvtudomány műhelyei [Workshops of present day Hungarian formal linguistics]*, 13–34. Budapest: Nemzeti Tankönyvkiadó.
- Laczkó, Tibor & György Rákosi. 2011. On particularly predicative particles in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '11 conference*, 299–319. Stanford, CA: CSLI Publications.
- Laczkó, Tibor & György Rákosi. 2013. Remarks on a novel LFG approach to spatial particle verb constructions in Hungarian. In Johan Brandtler, Valéria Molnár & Christer Platzak (eds.), *Approaches to Hungarian, volume 13: Papers from the 2011 Lund conference*, 149–177. Amsterdam: John Benjamins. DOI: [10.1075/atoh.13.08lac](https://doi.org/10.1075/atoh.13.08lac).
- Laczkó, Tibor & György Rákosi. 2019. Pronominal possessors and syntactic functions in the Hungarian possessive noun phrase. In Miriam Butt, Tracy Holloway King & Ida Toivonen (eds.), *Proceedings of the LFG '19 conference*, 149–169. Stanford, CA: CSLI Publications.
- Laczkó, Tibor & György Rákosi. 2008–2019. *HunGram: An XLE implementation*. Tech. rep. Debrecen: University of Debrecen.
- Laczkó, Tibor, György Rákosi & Péter Szűcs. 2020. On control and binding in Hungarian complex event nominals. In Miriam Butt & Ida Toivonen (eds.), *Proceedings of the LFG '20 conference*, 211–231. Stanford, CA: CSLI Publications.
- Lødrup, Helge. 2012. In search of a nominal COMP. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 383–404. Stanford, CA: CSLI Publications.
- Maticsák, Sándor. 2020. *A magyar nyelv eredete és rokonsága [The origin and the kinship of the Hungarian language]*. Budapest: Gondolat Kiadó.
- Miestamo, Matti, Anne Tamm & Beáta Wagner-Nagy (eds.). 2015. *Negation in Uralic languages*. Vol. 108 (Typological Studies in Language). Amsterdam: John Benjamins.
- Müller, Stefan. 2006. Phrasal or lexical constructions? *Language* 82. 850–883.
- Mycock, Louise. 2004. The wh-expletive construction. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*. Stanford, CA: CSLI Publications.
- Mycock, Louise. 2006. *The typology of wh-questions*. Manchester: University of Manchester. (Doctoral dissertation).

## 6 LFG and Finno-Ugric languages

- Mycock, Louise. 2010. Prominence in Hungarian: The prosody-syntax connection. *Transactions of the Philological Society* 108(3). 265–297. DOI: 10.1111/j.1467-968x.2010.01241.x.
- Mycock, Louise. 2013. Discourse functions of question words. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '13 conference*, 419–416. Stanford, CA: CSLI Publications.
- Nikolaeva, Irina. 2014. *A grammar of Tundra Nenets*. Berlin: Mouton de Gruyter. DOI: 10.1515/9783110320640.
- Nordlinger, Rachel & Louisa Sadler. 2004. Tense beyond the verb: Encoding clausal tense/aspect/mood on nominal dependents. *Natural Language & Linguistic Theory* 22(3). 597–641. DOI: 10.1023/b:nala.0000027720.41506.fe.
- Nordlinger, Rachel & Louisa Sadler. 2007. Verbless clauses: Revealing the structure within. 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*, 139–160. Stanford, CA: CSLI Publications.
- Parsons, Terence. 1990. *Events in the semantics of English: A study in subatomic semantics* (Current Studies in Linguistics 19). Cambridge, MA: The MIT Press.
- Patejuk, Agnieszka & Adam Przepiórkowski. 2016. Reducing grammatical functions in LFG. In Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King & Stefan Müller (eds.), *Proceedings of the joint 2016 conference on Head-Driven Phrase Structure Grammar and Lexical Functional Grammar*, 541–559. Stanford, CA: CSLI Publications.
- Payne, John & Erika Chisarik. 2000. Negation and focus in Hungarian: An optimality theory account. *Transactions of the Philological Society* 98. 185–230. DOI: 10.1111/1467-968x.00062.
- Pollard, Carl & Ivan A. Sag. 1987. *Information-based syntax and semantics*. Stanford, CA: CSLI Publications.
- Pylkkänen, Liina. 1997. The linking of event structure and grammatical functions in Finnish. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '97 conference*. Stanford, CA: CSLI Publications.
- Rákosi, György. 2003. Comitative arguments in Hungarian. In Willemijn Heeren, Dimitra Papangeli & Evangelia Vlachou (eds.), *Uil-OTS yearbook 2003*, 47–57. Utrecht Institute of Linguistics OTS.
- Rákosi, György. 2006a. *Dative experiencer predicates in Hungarian*. Published as volume 146 of the LOT series. Utrecht: University of Utrecht. (Doctoral dissertation).

*Tibor Laczkó*

- Rákosi, György. 2006b. On the need for a more refined approach to the argument-adjunct distinction: The case of dative experiencers in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '06 conference*. Stanford, CA: CSLI Publications.
- Rákosi, György. 2008. The inherently reflexive and the inherently reciprocal predicate in Hungarian: Each to their own argument structure. In Ekkehard König & Volker Gast (eds.), *Reciprocals and reflexives: Theoretical and typological explorations*, 411–450. Berlin: Mouton de Gruyter.
- Rákosi, György. 2009. Beyond identity: The case of a complex Hungarian reflexive. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 459–479. Stanford, CA: CSLI Publications.
- Rákosi, György. 2010. On snakes and locative binding in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 395–415. Stanford, CA: CSLI Publications.
- Rákosi, György. 2012. Non-core participant PPs are adjuncts. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 524–543. Stanford, CA: CSLI Publications.
- Rákosi, György & Tibor Laczkó. 2011. Inflecting spatial particles and shadows of the past in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '11 conference*, 440–460. Stanford, CA: CSLI Publications.
- Rätsep, Huno. 1969. Ühendverbide rektisioonistruktuuride iseärasustest eesti keeles [On the characteristic features of the government structures of complex verbs in Estonian]. *Emakeele Seltsi Aastaraamat* 14–15. 59–77.
- Rätsep, Huno. 1978. *Eesti keele lihtlausete tüübid* [Types of Estonian simple sentences]. Vol. 12 (ENSV TA Emakeele Seltsi Toimetised). Tallinn: Valgus. <https://dspace.ut.ee/handle/10062/28159>.
- Sahkai, Heete & Anne Tamm. 2018a. Estonian V2 is prosodic? Poster presentation at LFG '18 Conference, University of Vienna.
- Sahkai, Heete & Anne Tamm. 2018b. The syntax of contrastive topics in Estonian. In Marri Amon & Marie-Ange Julia (eds.), *Oralité, information, typologie / Orality, information, typology: hommage a M.M. Jocelyne Fernandez-Vest*, 399–418. Paris: L'Harmattan.
- Sahkai, Heete & Anne Tamm. 2019. Verb placement and accentuation: does prosody constrain the Estonian V2? *Open Linguistics* 5(1). 729–753. DOI: 10.1515/opli-2019-0040.
- Szabó, Ditta. 2017. *Evidencialitás az udmurt nyelvben* [Evidentiality in Udmurt]. Eötvös Loránd University, Budapest. (MA thesis).

- Szabó, Ditta. 2021. On evidentiality marking in Udmurt and its representation in LFG. In Kata Kubínyi, Judit Nagy & Anne Tamm (eds.), *In memoriam of Anne Vainikka. Conference contributions II, November 22-23, 2019*, 60–97. Budapest: Károli Gáspár University.
- Szabolcsi, Anna. 1994. The noun phrase. In Ferenc Kiefer & Katalin É. Kiss (eds.), *The syntactic structure of Hungarian*, vol. 27 (Syntax and Semantics), 179–274. Academic Press.
- Szűcs, Péter. 2013. A fókuszemelésről új adatok tükrében [On focus raising in the light of new data]. In Zsuzsanna Gécseg (ed.), *Lingdok 13: Nyelvészdoktoranduszok dolgozatai*, 257–278. Szeged: Szegedi Tudományegyetem Nyelvtudományi Doktori Iskola.
- Szűcs, Péter. 2014. A magyar „operátoremelés” mint prolepszis [Hungarian operator raising as prolepsis]. In Zsuzsanna Gécseg (ed.), *Lingdok 14: Nyelvészdoktoranduszok dolgozatai*, 185–204. Szeged: Szegedi Tudományegyetem Nyelvtudományi Doktori Iskola.
- Szűcs, Péter. 2018a. A COMP-less approach to Hungarian complement clauses. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '18 conference*, 325–342. Stanford, CA: CSLI Publications.
- Szűcs, Péter. 2018b. Operator fronting in Hungarian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '18 conference*, 343–363. Stanford, CA: CSLI Publications.
- Szűcs, Péter. 2019. Remarks on binding and control data in Hungarian complex event nominals. *Argumentum* 15. 650–664.
- Tael, Kaja. 1988. *Sõnajäremallid eesti keeles (võrrelduna soome keelega)* [Word order patterns in Estonian (compared to Finnish)]. Tallinn: Teaduste Akadeemia Keele ja Kirjanduse Instituut.
- Tamm, Anne. 2004a. Eesti ja ungari keele verbiaspekti modelleerimise probleeme [Problems of modeling Estonian and Hungarian verbal aspect]. In Marju Ilves & János Pusztay (eds.), *Folia estonica tomus xi. Ész-t-magyar összevetés [Estonian-Hungarian comparison]*, 128–147. Szombathely: Savaria University Press.
- Tamm, Anne. 2004b. On the grammaticalization of the Estonian perfective particles. *Acta Linguistica Hungarica* 51(1-2). 143–169. DOI: 10.1556/aling.51.2004.1-2.6.
- Tamm, Anne. 2004c. *Relations between Estonian aspect, verbs, and case*. Budapest: Eötvös Loránd University. (Doctoral dissertation).
- Tamm, Anne. 2006. Estonian transitive verbs and object case. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '06 conference*. Stanford, CA: CSLI Publications.

Tibor Laczkó

- Tamm, Anne. 2007a. Aspect and the Estonian partitive objects: A review of arguments for analysing partitive NPs as instances of incorporation. In Daniele Monticelli & Anu Treikelder (eds.), *Aspect in languages and theories: Similarities and differences. Studia romanica tartuensia* 6, 205–26. Tartu: University of Tartu Press.
- Tamm, Anne. 2007b. Estonian object and adverbial case with verbs of motion. In Márta Csepregi & Virpi Masonen (eds.), *Grammatika és kontextus [Grammar and context]* (Urálsztikai tanulmányok 17 [Uralistic Studies 17]), 319–330. Budapest: ELTE BTK Finnugor Tanszék.
- Tamm, Anne. 2007c. Representing achievements from Estonian transitive sentences. In Magnus Sahlgren & Ola Knutsson (eds.), *Proceedings of the workshop ‘Semantic Content Acquisition and Representation’ (SCAR)*, 28–35.
- Tamm, Anne. 2008. Partitive morphosemantics across Estonian grammatical categories, and case variation with equi and raising. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’08 conference*, 473–493. Stanford, CA: CSLI Publications.
- Tamm, Anne. 2011a. Cross-categorical spatial case in the Finnic non-finite system: Focus on the absentive TAM semantics and pragmatics of the Estonian inessive m-formative non-finites. *Linguistics: An Interdisciplinary Journal of the Language Sciences* 49(4). 835–944. DOI: 10.1515/ling.2011.025.
- Tamm, Anne. 2011b. The case denoting “within” denotes locative relationships with more noun-like non-finites and aspectual relationships in more verb-like non-finites in Finnic. In Sándor Csúcs, Nóra Falk, Viktória Tóth & Gábor Zaicz (eds.), *Proceedings of Congressus XI Internationalis Fenno-Ugristarum*, 6, 242–249. Piliscsaba: Reguly Társaság.
- Tamm, Anne. 2012a. Finniségi tárgyeset mint a kategóriákat áthidaló eset egyik altípusa: Az argumentumon megjelenő aspektusjelölő toldalék [Accusative in Finnic as a subtype of cases across categories: The aspect marker appearing on arguments]. In Gábor Alberti, Judit Kleiber & Judit Farkas (eds.), *Vonzásban és változásban [In attraction and change]*, 181–207. Pécs: Pécsi Tudományegyetem Nyelvtudományi Doktori Iskola.
- Tamm, Anne. 2012b. Intermoduláris megközelítések: Az észti ige, aspektus és eset kölcsönhatásai [Intermodular approaches: The interactions of the verb, aspect and case in Estonian]. In Kristiina Lutsar & János Pusztay (eds.), *Észt-magyar összevetés VI [Estonian-Hungarian comparison VI]. Folia Estonica*, 25–149. Szombathely: Savaria University Press.
- Tamm, Anne. 2012c. *Scalar verb classes: Scalarity, thematic roles, and arguments in the Estonian aspectual lexicon*. Florence: University of Florence Press. DOI: 10.36253/978-88-6655-055-6.



- Tamm, Anne & Anne Vainikka. 2018. An overview of generative works on Finnish and Estonian syntax. *Finno-Ugric Languages and Linguistics* 7(2). 80–89.
- Tamm, Anne & Anne Vainikka (eds.). to appear. Cambridge: Cambridge University Press.
- Toivonen, Ida. 2000. The morphosyntax of Finnish possessives. *Natural Language & Linguistic Theory* 18(3). 579–609.
- Toivonen, Ida. 2001. Language change, lexical features and Finnish possessors. In Miriam Butt & Tracy Holloway King (eds.), *Time over matter: Diachronic perspectives on morphosyntax*, 209–225. Stanford, CA: CSLI Publications.
- Toivonen, Ida. 2007. Verbal agreement in Inari Saami. In Ida Toivonen & Diane Nelson (eds.), *Saami linguistics* (Current Issues in Linguistic Theory 288), 227–258. Amsterdam: John Benjamins. DOI: 10.1075/cilt.288.09toi.
- Torn, Reeli. 2006. Oblique dependents in Estonian: An LFG perspective. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '06 conference*. Stanford, CA: CSLI Publications.
- Vainikka, Anne & Joan Maling. 1996. Is partitive case inherent or structural? In Jack Hoeksema (ed.), *Partitives: Studies on the syntax and semantics of partitive and related constructions*, 179–208. Amsterdam: Mouton de Gruyter. DOI: 10.1515/9783110908985.179.
- Vilkuna, Maria. 1995. Discourse configurationality in Finnish. In Katalin É. Kiss (ed.), *Discourse configurational languages*, 244–268. Oxford: Oxford University Press.
- Viszket, Anita. 2004. *Argumentumstruktúra és lexikon. A pusztá NP grammatikai sajátosságai a magyarban és ezek következményei a predikátumok lexikonbeli argumentumstruktúrájára* [Argument structure and the lexicon. The grammatical properties of the bare NP in Hungarian and their impact on the argument structures of predicates in the lexicon]. Budapest: Eötvös Loránd University. (Doctoral dissertation).
- Volkova, Anna A. 2014. *Licensing reflexivity: Unity and variation among selected Uralic languages*. Utrecht: University of Utrecht. (Doctoral dissertation).
- Volkova, Anna A. to appear. Reflexivity and binding in Uralic languages. In Anne Tamm & Anne Vainikka (eds.), *Uralic syntax*. Cambridge: Cambridge University Press.
- Zsirai, Miklós. 1933. Az obi-ugor igekötők [Ob-Ugric verbal particles]. *Értekezések a Nyelv- és Széptudományi Osztály Köréből* 25(3). 41–82.





# Chapter 7

## LFG and Romance languages

Alex Alsina

Universitat Pompeu Fabra

This chapter is an overview of the main topics in the Romance languages that have been the object of analysis within LFG. The topics reviewed include the analysis of verbal clitics, considering their morphological and c-structure status, their role in f-structure, and the role of the anaphoric reflexive clitics in a-structure, the grammatical function of direct and indirect objects and of clausal complements, passive and impersonal constructions, and complex predicates such as the causative construction.

### 1 Introduction

This section consists of a brief presentation of the Romance languages and an overview of the chapter.

#### 1.1 Brief presentation of the Romance languages

The Romance languages developed out of the varieties of Latin spoken in the areas under Roman domination as a result of the expansion of Latin throughout the territories around the Mediterranean Sea from the fifth century BC to the sixth century AD. The main present day Romance languages with a standard form and/or official status in some state or region within a state are:

- The closely related Portuguese and Galician;
- Spanish, or Castilian;
- Catalan, with Valencian as a regional name;



*Alex Alsina*

- French;
- Occitan, with a variety of regional names including Provençal, Langue d'Oc, Gascon, Limousin, etc.;
- Sardinian;
- Italian;
- Raeto-Romance, with Romansh, Ladin and Friulian as regional names;
- Romanian.

In addition, there are a number of languages without an official status, such as Asturian and Aragonese in Spain, Walloon, Picard and Bourguignon in France, the Italo-Romance varieties Piedmontese, Ligurian, Lombard, Sicilian, Neapolitan in Italy, and Corsican in France, and the Daco-Romance varieties Aromanian, Istro-Romanian, and Megleno-Romanian, to name just a few. As a consequence of the colonial policies of European states from the fifteenth to the nineteenth centuries, some of the Romance languages have large numbers of speakers outside Europe; this is the case of Spanish, Portuguese, and French, which are, in this order, the Romance languages with the largest numbers of speakers.

Because of their (relatively recent) common ancestry, the Romance languages share many structural patterns, but they also have significant differences. Readers interested in finding more information about any aspect of this language family should consult Ledgeway & Maiden (2016).

## 1.2 Overview of the chapter

The choice of topics dealt with in this chapter is conditioned by the existence of LFG work on specific topics, theoretical interest, and space limitations. Most of the LFG work on Romance is on Spanish, French, Italian, Catalan, and Portuguese. Consequently, this chapter will deal mostly with these languages.

Section 2 focuses on so-called clitics in Romance. First, it addresses the debate about their morphological status: are clitics affixes or independent words? Second, it addresses their syntactic status: do they fill a grammatical function (GF) and, if so, what GFs can they fill? Can they be agreement markers? Do they have other roles? And, finally, the status of the anaphoric reflexive clitic is debated. Section 3 discusses arguments, GFs, and case and addresses issues such as the inventory of GFs in LFG and what GFs should be used for objects in the

7 *LFG and Romance languages*

Romance languages, subject-object alternations, passivization, etc. Section 4 discusses complex predicate constructions such as the causative construction and the restructuring construction.

Following are some topics not discussed in detail in this chapter:

- The phenomena that, in Rizzi (1997) and subsequent work, are known as corresponding to the structure of the left periphery. Although there is not much LFG work on this topic within the Romance languages, Estigarribia (2005, 2013) analyses clitic left dislocation in Spanish; Gazdik (2008, 2010) studies interrogatives and multiple questions in French; and Zipf & Quaglia (2017) discuss word order and information structure in Italian matrix wh-questions. See **chapters/InformationStructure** for a general discussion.
- Determiners and the structure of the NP. A salient feature of the Romance languages in general is the existence of a clitic-like definite article. In many of these languages it is homophonous (or partially so) with the third person pronominal accusative clitic. Article-preposition contracted forms in French are analyzed in Wescoat (2007) as lexical items involving lexical sharing. Alsina (2010) takes this idea a step further and assumes that the definite article in Catalan is always an affix that attaches to a word with lexical sharing. Alsina (2011) identifies three types of determiners in Catalan depending on whether they must co-occur with a head noun, may (but need not) co-occur with a head noun, or cannot do so, and provides an analysis within LFG.
- Agreement. Verb agreement is generally taken to be agreement with the subject. However, Alsina & Vigo (2014) show that the finite verb may agree with a non-subject (a nominative complement) in Catalan and Spanish, as can be seen in copular constructions; Alsina & Yang (2018) extend this assumption to intransitive clauses with an indefinite postverbal logical subject. Carretero García (2017) proposes an analysis of the special form of adjectives used for agreement with non-count nouns in Asturian.
- Diachrony. The diachronic development of infinitival complements from Latin to the Romance languages is the topic of Vincent (2019).
- Finiteness and tense, in connection with the morphology-syntax interface, are dealt with in Barron (2000) and Schwarze (2001a), using data from Italian and French, respectively.

*Alex Alsina*

- Auxiliaries. Butt et al. (1996) develop the idea of having a level of representation different from f-structure, called m-structure, for the analysis of auxiliaries in French, as well as in English and German. Schwarze (1996) proposes an analysis of auxiliaries in Spanish, Italian and French, including auxiliary selection in the latter two languages.

## 2 Clitics

The term “clitic” in this chapter is used as a purely descriptive term (without any theoretical implications) to refer to the class of phonologically dependent particles that attach to a verb and generally provide information about a GF of the clause. Section 2.1 focuses on the debate as to whether clitics are syntactically independent words (though phonologically dependent) or affixes, what their correct analysis should be, and what implications this analysis has. In Section 2.2, we examine the f-structure status of clitics as the expression of an argument, as an agreement marker, and as the expression of a non-argument of the verb. The reflexive clitic, in its “anaphoric” use, is discussed in Section 2.3.

### 2.1 Morphological status

One of the central issues in the analysis of clitics is their morphological status: Are they independent syntactic constituents or are they affixes? If they are affixes, should we treat them as morphemes – linguistic signs consisting of a phonological representation and a semantic and f-structure representation – or should we treat them as the overt realization of particular bundles of morphological or syntactic features within a realizational approach to morphology?

#### 2.1.1 Affixes vs. independent words

The most common assumption in connection with this issue in LFG is that they are independent syntactic constituents. This is not only the oldest approach, as it is found in the earliest analyses of clitics within LFG, as in Grimshaw (1982), but it is a very prevalent one, as it is found up to the present (see, for example, Schwarze (2001b) for French and Italian, Estigarribia (2005, 2013) for Spanish, Quaglia 2012 for Italian, and Barbu & Toivonen 2018 for Romanian). Grimshaw (1982: 90) posits the following c-structure rule in order to account for the position of clitics in French:

$$(1) \quad V' \longrightarrow (CL)_1 (CL)_2 (CL)_3 (AUX) V$$

In this approach, clitics are a special grammatical category (CL) that occupies a position in the c-structure. By a rule such as (1), the position of clitics is restricted to being adjacent to a verb (or an auxiliary).

However, proponents of clitics as syntactic constituents generally do not present arguments in favor of their position and against treating clitics as affixes, presumably because that position is seen as the default assumption, given that the standard orthographies in general separate clitics from their hosts by means of spaces, hyphens or apostrophes, at least in preverbal position, which induces the belief that clitics are words. (But see Schwarze 2001b, who makes an explicit defense of clitics as c-structure constituents, in French and Italian.)

On the other hand, proponents of treating clitics as affixes have presented evidence in favor of this assumption that is highly problematic for the assumption that clitics are independent syntactic constituents. Evidence that clitics are morphological units (not c-structure constituents) has been presented, within different frameworks, by Bonet (1991, 1995); Miller (1992); Crysmann (1997); Miller & Sag (1997); Monachesi (1999); Luís & Sadler (2003); Luís & Spencer (2005), among others. Some of the evidence, of a strictly syntactic nature, is that clitics cannot be topicalized, cannot be substituted by full pronouns, cannot be coordinated, and cannot be modified. The following Portuguese examples illustrate the failure of coordination of clitics:<sup>1</sup>

(2) Portuguese (Crysmann 1997)

- a. \*eu vi            o            e    Paulo  
    I    saw.1SG 3SG.M.ACC and Paul  
    'I saw him and Paul.'
- b. \*eu não o            e    a            conheço  
    I    not 3SG.M.ACC and 3SG.F.ACC know.1SG  
    'I do not know him and her.'

There is also evidence that can be classified as morphophonological. Clitics exhibit a high degree of selection with respect to their host: in most Romance languages, the clitic cluster must be adjacent to the verb. This is always the case

---

<sup>1</sup>Examples, including cited ones, are glossed according to the Leipzig glossing rules, replacing the original glosses, if necessary. Unreferenced examples reflect the author's judgments. Clitics are glossed indicating only the corresponding features of person, number, gender, and case that are morphologically relevant. The reflexive clitic (*se*, *si*, *s'*, and cognate forms) is glossed as REFL (even when its meaning is not reflexive). Forms that cannot be glossed in a simple way are glossed with the form in small caps; example: the genitive and partitive clitic *en* in Catalan or French is glossed as EN.

Alex Alsina

when the clitic cluster is postverbal. The exception to the adjacency requirement of the verb and the clitic cluster is only found when the clitic cluster is preverbal in modern Portuguese (Crysmann 1997; Luís & Otaguro 2004), as well as in medieval Spanish and Portuguese (Fontana 1993, 1996; Fischer 2002), and only in very restricted contexts. There are morphophonological alternations that are restricted to clitic combinations. For example, in Portuguese, when a third person accusative clitic (-o, -a, -os, -as) is in a clitic combination (either with a verb or with another clitic) following an oral coronal continuant (/s/, /z/, /r/), this consonant is replaced by [l] (see Crysmann 1997): *comprar + o* → *comprá-lo* buy.INF it, *nos o dão* → *no-lo dão* us-it give.3PL, etc. This alternation does not occur across word boundaries: *todos os alunos* ‘all the students’. The same clitics appear preceded by /n/ when suffixed to a verb form ending in a nasal vowel, as in *eles conhecem + o/a* → *eles conhecem-no/na* ‘they know him/her’ (see Crysmann 1997), but this nasal insertion does not occur across word boundaries: *eles conhecem o aluno/a aluna* vs. *\*eles conhecem no aluno/na aluna* ‘they know the student.M/the student.F’.

One of the most compelling sources of morphophonological evidence for the affixal nature of clitics is the existence of opaque clitic combinations, i.e. combinations of clitics that do not coincide with the form of the corresponding clitics used in isolation (Bonet 1995). One of the clearest examples of opaque clitic combinations is the so-called “spurious *se*” in Spanish. While the clitic form of the third person singular indirect object is *le* in isolation, as in (3b), when it combines with a third person accusative object, such as *lo* in (3a), it adopts the form *se*, as in (3c), elsewhere used only as a third person reflexive clitic. The transparent combination *\*le lo* (or *\*lo le*) does not exist.

(3) Spanish (Bonet 1995: 608)

- a. El premio, lo                      dieron    a Pedro ayer.  
the price    3SG.M.ACC gave.3PL to Pedro yesterday  
‘The price, they gave it to Pedro yesterday.’
- b. A Pedro, le                      dieron    el premio ayer.  
to Pedro, 3SG.DAT gave.3PL the price    yesterday  
‘Pedro, they gave him the price yesterday.’
- c. A Pedro, el premio, se lo                      dieron    ayer.  
A Pedro the price    SE 3SG.M.ACC gave.3PL yesterday  
‘Pedro, the price, they gave it to him yesterday.’

Another instance of an opaque clitic combination, among those reported in Bonet (1995), is the combination in standard Italian of the impersonal clitic *si* with the

third person reflexive clitic *si*: instead of the expected *si si* sequence (possible in certain dialects of Italian), the sequence *ci si* is found:

- (4) Italian (Bonet 1995: 609)
- a. Lo            si            sveglia.  
3SG.M.ACC IMPERS wake.up.3SG  
‘one wakes him/it up’
  - b. Se lo compra.  
REFL 3SG.M.ACC buy.3SG  
‘S/he buys it for herself/himself’
  - c. Ci si lava  
REFL IMPERS wash.3SG  
‘one washes oneself’

These opaque clitic combinations are completely unexpected under the treatment of clitics as words and very difficult to explain in that approach. On the other hand, if clitics are affixes, this kind of allomorphy is much more natural.

In addition, there is phonological evidence for the affixal status of clitics. One of the sources of such evidence is word stress. While clitics in most cases are stressless and have no effect on the stress pattern of the word they are attached to, there are some Romance varieties in which clitics affect the stress pattern of their host. This is the situation in the Catalan dialects of Mallorca and Minorca: the first column in (5) illustrates verb forms without postverbal clitics and the second column shows the same verb forms with postverbal clitics:

- (5) Mallorcan Catalan (Colomina i Castanyer 2002: 579)
- |                          |                                  |
|--------------------------|----------------------------------|
| dona ‘give’ [ˈdonə]      | dona’m ‘give me’ [doˈnəm]        |
| agafa ‘pick up’ [əˈɣafə] | agafa’l ‘pick it up’ [əɣəˈfəl]   |
| entra ‘enter’ [ˈentrə]   | entra-hi ‘enter there’ [ənˈtrəj] |

In some dialects, the presence of a clitic in postverbal position causes stress to be placed on the final syllable of the verb form, instead of on the penultimate syllable. Given that word stress in Catalan does not depend on elements external to the word, one must conclude that clitics are part of the word at the point in the derivation in which word stress placement rules apply. In other dialects, clitics are affixes that do not affect stress placement.

All of these facts argue conclusively for treating clitics as word parts, specifically, affixes.

Alex Alsina

### 2.1.2 Alternative analyses of clitics as affixes

The affixal status of clitics is consistent with two approaches to morphology, in particular, to inflectional morphology: the morpheme-based approach and the realizational approach. Within the morpheme-based approach, an affix is a linguistic sign, consisting of a phonological representation and a semantic and/or syntactic representation. Under this view, an affix is very much like a word, only, instead of being an element that can appear as a terminal node in the c-structure, it is part of a word and combines with other word parts in a tree structure to form a word. Within the realizational approach, an affix is the phonological realization or spell-out of semantic and/or syntactic features of a word, possibly mediated by morphological features. In inflectional morphology, for every lexeme there are as many feature combinations as there are possible forms in the paradigm of the lexeme and there are rules spelling out specific features or feature combinations as particular affixes.<sup>2</sup>

To compare the two approaches, consider the form *lo* that appears in Italian examples such as (4a),(4b). In a morpheme-based approach, the form *lo* would have a dictionary entry that, instead of specifying its grammatical category, as it would for a word, indicates what kind of word part it is.<sup>3</sup> For the purpose of illustration, we can assume that form would be classified as a special kind of affix, which we can call *cl* (for clitic), as in (6):

- (6) *lo*                      *cl*    (↑ PRED)=‘PRO’  
    (↑ CASE)=ACC  
    (↑ PERS)=3  
    (↑ GEND)=M  
    (↑ NUM)=SG

Affixes of type *cl* combine with other *cl* elements to form a clitic cluster (CCL). In Italian, a CCL attaches preverbally (as a prefix) to finite verb forms except for imperatives and postverbally (as a suffix) to imperatives and non-finite verb forms.

<sup>2</sup>These two approaches are mutually exclusive within LFG, although they are not necessarily so in a derivational framework such as Minimalism. As pointed out by a reviewer, this would be the case with Distributed Morphology (Halle & Marantz 1993), which is claimed to be both morpheme-based and realizational. In this framework, morphemes are bundles of morphosyntactic features as terminal nodes in the syntax without a phonological representation and are subsequently assigned a phonological form.

<sup>3</sup>Standardly, in LFG the term “lexical entry” refers to the information associated with a fully inflected word, as it appears in the syntax. Since clitics, as affixes, are not fully inflected words, I avoid using the term “lexical entry” to refer to the phonological, morphological, f-structure, and semantic information that characterizes a sublexical element such as an affix, but use the term “dictionary entry” instead.



There has to be some mechanism to place clitics in the right order within a CCL: As we see in (4), *lo* precedes the impersonal clitic *si* but follows the third person reflexive *si* (which takes on the form *se* in this context). This can be achieved by having a template with several clitic positions and having each clitic subclassified as to the position it occupies in the template. Alternatively, there can be linear precedence rules that are sensitive to the syntactic features of the clitics (such as person, case, reflexivity, etc.) and order the clitics within a CCL according to these features.

In a realizational approach, there are rules that spell out bundles of f-structure features (or the corresponding morphological features) as the appropriate clitic form. So, if a verb form has the f-structure features in (6), a rule is triggered that introduces the form *lo* in a CCL, along the following lines:

- (7) [PRED 'PRO', CASE ACC, PERS 3, GEND M, NUM SG]  $\rightarrow$  CL {...lo...}

As in the previous approach, there would also have to be a mechanism such as a template, linear precedence rules, or ordered blocks of rules to obtain the right order of clitics when more than one is present in a CCL.

At first sight there might seem to be little difference between the two approaches. Both approaches can account with a similar degree of success for the strictly syntactic evidence for the affixal nature of clitics noted in Section 2.1.1 (such as the failure to be topicalized, substituted by full pronouns, coordinated or modified): these processes affect c-structure units, which clitics are not in either approach. The phonological evidence for the affixal status of clitics (e.g. instances in which stress assignment applies to the word structure that includes clitics) is accounted for in a similar way whether affixes are viewed as morphemes or as the product of spell-out rules.

Many of the morphophonological arguments for the affixal status of clitics can also be accounted for in either approach. To account for an allomorphic alternation such as the *o/lo/no* alternation in Portuguese noted earlier, within the morpheme-based approach, we would have to assume that the third person singular accusative masculine clitic morpheme has three allomorphs that are phonologically conditioned; within the realizational approach, we would have to assume that there are three different spell-out rules for the same syntactic (or morphological) feature bundle each one with a different phonological context. The existence of opaque clitic combinations, such as the ones illustrated in (3) and (4), is probably the strongest argument in favor of the realizational approach. From the morphemic perspective, these can be thought of simply as instances of allomorphic alternations. To use the example of the Spanish spurious *se*, illustrated in (3), the third person dative clitic morpheme would have two

Alex Alsina

allomorphs: *se*, when it co-occurs with another third person clitic; and *le*, elsewhere. A problem with this approach is that it fails to explain the observation by Bonet (1995) that, in opaque clitic combinations, the unexpected form always coincides with a clitic that exists independently in the language. If the third person dative *se* is an allomorph of the more general *le*, it is just an accident that it is homophonous with the third person reflexive *se*; it could just as easily be *che*, *je*, *na*, or any other form that does not coincide with an existing clitic. On the other hand, the realizational approach has the means of capturing that observation, as in Bonet (1995); see also Grimshaw (1997) using OT.<sup>4</sup>

### 2.1.3 Proclisis and enclisis in European Portuguese

This subsection illustrates to what extent Romance data can call standard LFG assumptions into question, in particular, the way the Lexical Integrity Principle is to be interpreted. The position of clitics (or, more exactly, of the CCL) in European Portuguese (EP) with respect to the verb of their clause poses an important problem for theories of syntax, morphology, and the syntax-morphology interface. Two properties that distinguish EP from the other modern Romance languages are relevant in this context:<sup>5</sup>

- With finite verb forms, the CCL can appear after the verb (enclisis) or before it (proclisis), depending on the kind of syntactic constituent, if any, that precedes the verb.
- When it appears before the verb, it need not be adjacent to it, but may be separated from it by words such as some adverbs and the negation *não* (interpolation) (see Section 2.1.1).

These properties are a problem for the affixal treatment of clitics. If we assume that clitics are affixes in both preverbal and postverbal position, the fact that the choice between the two positions is dependent on a syntactic property (the presence or absence of certain types of syntactic constituents before the verb)

<sup>4</sup>The facts involving the expression and omission of the reflexive clitic in Catalan presented in Alsina (2020) are further evidence for the realizational treatment of clitics in Romance.

<sup>5</sup>An additional specificity of the positioning of CCL in EP is the phenomenon of mesocclisis: With future and conditional verb forms, the enclitic position is not after the tense, aspect and person affixes, but before them. See Luís & Spencer (2005) for an analysis within LFG and realizational morphology. When a present tense form such as *mostramos* ‘we show’ combines with the clitic complex *lho* (3.DAT+3SG.M.ACC) in enclitic position, the result is *mostramos-lho* ‘we show it to him’, but, if instead we use a future tense form such as *mostraremos* ‘we will show’, the enclitic attachment of *lho* results in *mostrar-lho-emos*, not \**mostraremos-lho*.

is a *prima facie* problem for that assumption. The standard view of the syntax-morphology interface in a lexicalist framework assumes that the morphology may impose constraints on the syntax, but the syntax cannot impose constraints on the morphology. But, in the case in point, a particular morphological property – the linearization of CCL before or after V – is determined by the syntax.

Example (8) shows that the same finite verb form, here *vê*, can take a clitic before it, as in *me vê*, or after it, as in *vê-me*, depending on what precedes it.

- (8) Portuguese (Luís & Otaguro 2004)
- a. O João raramente me vê.  
the João rarely 1SG see.3SG
  - b. O João vê-me raramente.  
the João see.3SG-1SG rarely  
'João rarely sees me.'

The accepted assumption in work such as Luís & Sadler (2003), Luís & Otaguro (2004, 2005), and Luís & Spencer (2005), among others, is that enclisis is the default linearization of CCL and the verb in EP, whereas proclisis is triggered by the presence of certain c-structure constituents in preverbal position, which can be referred to as proclisis-triggers. So, for example, a non-quantified preverbal subject, such as *o João* in (8), is not a proclisis-trigger, which implies that the default option of enclisis is chosen in (8b); on the other hand, the adverb *raramente* in preverbal position is a proclisis-trigger, which explains the proclitic sequence *me vê* in (8a).

The approach adopted in Luís & Sadler (2003) is that all syntactic constituents that are proclisis-triggers are associated with the f-structure feature ( $\uparrow$  TYPE)=NON-NEUTRAL (or with the morphological feature [Restricted:Yes] in Luís & Otaguro 2004). For example, the negative element *não* is associated with this feature. (It has not been possible to find a common configurational or semantic/discourse denominator for the set of syntactic contexts that trigger proclisis; hence the proposal of having an f-structure feature for proclisis.)

The linearization rule 'Proclitic-LR', which ensures that CCL is placed preverbally, applies only under the existence of the ( $\uparrow$  TYPE)=NON-NEUTRAL feature in the f-structure of the verb. In the absence of this feature, the linearization rule that places CCL postverbally applies. So, the TYPE feature reflects the idea that proclisis is the marked option in EP.

However, the two alternative sequences *vê-me* and *me vê* are not identical from the syntactic point of view: even though they are both assumed to be a word from the morphological point of view, the form with enclisis, *vê-me*, is assumed

Alex Alsina

to constitute a single  $X^0$  (either I or V), whereas the form with proclisis, *me vê*, is assumed to correspond to two different c-structure positions. This assumption is necessary in order to account for two phenomena: scope over coordinated VPs and so-called interpolation. Focusing on interpolation, we find that certain words, which are clearly independent syntactic constituents, can appear between the proclitic CCL and the verb. These words can be the negative element *não*, certain adverbs like *ainda* ‘yet’, subject pronominals, and a combination of them, as in (9):

- (9) Portuguese (Luís & Otoguro 2005)  
 ... acho        que ela o            ainda não disse.  
 ... think.1SG that she 3SG.M.ACC yet    not told.3SG  
 ‘... I think that s/he hasn’t told it to him/her/them yet.’

The clitic *o* in (9) is separated from the verb *disse* by two words: *ainda* and *não*. This indicates that *o* and *disse* must be two independent c-structure elements (c-structure words). On the other hand, according to Luís & Sadler (2003), Luís & Otoguro (2004, 2005), and Luís & Spencer (2005), these two elements constitute a single unit at the morphological level (a morphological word). This is a departure from the standard idea in LFG that words – the minimal units of c-structure – are the output of the morphological component and, so, there should be no reason to distinguish between a c-structure word and a morphological word.

The exact implementation of the syntactic representation of the form with proclisis varies depending on the work. In Luís & Sadler (2003), the proclitic CCL attaches to the left of the VP headed by the verb that constitutes a morphological word with the preverbal CCL. In Luís & Otoguro (2004), it is assumed that, in certain cases, i.e. proclisis, a morphological token may correspond to two or more c-structure terminals.

In both approaches, a morphological unit is decomposed into two elements in the c-structure, which is a clear violation of the Lexical Integrity Principle – the idea that the internal structure of words is invisible to the c-structure. More specifically, this treatment of proclisis in Portuguese can be seen as a violation of Zwicky’s Principle of Morphology-Free Syntax, according to which “syntactic rules cannot make reference to the internal morphological composition of words or to particular rules involved in their morphological derivation” (Zwicky 1987: 650), which he considers equivalent to the Lexicalist Hypothesis or the belief that syntax is blind to morphology (O’Neill 2016: 244).

## 2.2 F-structure status

In this subsection we address the issue of the GF that the clitic corresponds to, if any, and its status as a pronoun or an agreement marker, leaving aside the reflexive clitic, to be discussed in Section 2.3 and Section 3.2.

### 2.2.1 The GF the clitic corresponds to

In most cases, a clitic corresponds to a GF in its clause. In some languages (e.g., Spanish, Portuguese, Catalan, Italian), a clitic cannot correspond to the subject; it can correspond to an object only, both accusative and dative, as in Spanish and Portuguese, or to an object, as well as an oblique, as in Catalan and Italian. In French, clitics can correspond to a subject, in addition to objects and obliques.<sup>6</sup>

The most common situation with clitics is that in which the clitic is in complementary distribution with the phrasal expression of the GF that the clitic corresponds to. As Grimshaw (1982: 88) notes for French, “accusative clitics are in complementary distribution with NP objects.” With a verb like *voit* ‘sees’, which requires a direct object, either an NP object or an accusative clitic satisfies this requirement, as in (10b) and (10c) respectively, but they cannot co-occur, as in (10d):

(10) French (Grimshaw 1982: 88)

- a. \*Jean voit.  
‘John sees.’
- b. Jean voit l’homme.  
‘John sees the man.’
- c. Jean le voit.  
John 3SG.M.ACC see.3SG  
‘John sees him.’
- d. \*Jean le voit l’homme.  
John 3SG.M.ACC see.3SG the.man  
‘John sees him the man.’

How these facts are explained depends in part on whether we treat clitics as c-structure constituents or as affixes and, within the affixal treatment, as morphemes or as exponents of morphological or syntactic features. We shall consider

<sup>6</sup>Some Northern Italian languages also require a subject clitic, in a wide range of modalities (see Renzi & Vanelli 1983 and Cardinaletti & Repetti 2010). See Poletto & Tortora (2016) for variation in subject clitics in the different Romance languages that are claimed to have subject clitics.

*Alex Alsina*

the different approaches to clitics in explaining distributional facts such as those illustrated in (10).

Grimshaw (1982) takes the position that clitics are c-structure constituents belonging to the CL (clitic) grammatical category – let’s call it the clitic-as-word approach. The observation that the direct object requirement is satisfied by either an NP following the verb or an accusative clitic before the verb is explained: (a) by annotating as an OBJ both the NP daughter of VP and one of the CL (clitic) positions daughters of V’; (b) by assuming that both nouns and pronouns, including pronominal clitics, have a PRED feature in their lexical entries and that the clitic *le* has the lexical entry in (11); and (c) by appealing to the standard well-formedness conditions of Consistency, Completeness, and Coherence.

- (11) *le* CL (↑ PRED) = 'PRO'  
 (↑ CASE) = ACC  
 (↑ NUM) = SG  
 (↑ PERS) = 3  
 (↑ GEND) = M

This clitic, in the appropriate CL position, satisfies the OBJ requirement of the verb and provides the necessary PRED feature to satisfy Completeness. It is an alternative to the NP realization of the object, in which a noun provides the PRED feature. This explains the alternative expression of the object illustrated in (10b,10c). In addition, if both ways of expressing the OBJ are used in the same clause, a violation of Consistency results, as the OBJ would have two PRED values, given the convention that PRED values are not unifiable, which accounts for (10d).

The affixal treatment of clitics is common to the two approaches in Section 2.1.2. In the morpheme-based approach, the main difference with the clitic-as-word approach is that clitics are not joined to a verb in the c-structure, but are joined to it in the lexicon. A clitic such as *le*, being a morpheme, has a “sublexical” entry identical or very similar to that in (11), except that “CL” is not a c-structure category, but a type of affix. One could either assume that there is a word template with different clitic affix positions, one of which would be annotated as the OBJ, and then a sublexical entry like that in (11) would fit into that position providing the f-structure features to the OBJ. Alternatively, the clitic-affix *le* would be specified in its sublexical entry with features indicating the GF they correspond to, such as (↑ OBJ PRED) = ‘PRO’, (↑ OBJ CASE) = ACC, etc. The concatenation of a clitic (or clitic cluster) with a verb yields a word whose f-structure information is the union of that of the clitic and that of the verb, so that a word such as *le voit*

carries the f-structure information of the clitic and the f-structure information of the verb *voit*. This would be an instance of pronominal incorporation similar to the analysis of object markers in Chicheŵa by Bresnan & Mchombo (1987).

Within the realizational approach, the phonological representation of the clitic – *le*, in the French example (10b) – is the result of exponence rules and linearization rules, to use the concepts of Luís & Sadler (2003). Adapting the approach of Luís & Sadler (2003) to the present example, we can assume that one of the forms of the paradigm of the verb *voit* has the morphological feature (or m-feature) bundle {ACC,3,SG,M}. This feature bundle is realized phonologically as *le* and this exponent is linearized preceding the verb stem, giving the form *le voit*. In addition, there is a mapping between the m-features and f-structure features. Specifically, the m-feature bundle {ACC,3,SG,M} corresponds to the same f-structure features of the OBJ as those in (11). This is shown schematically in (12a) for the phonological realization and in (12b) for the f-structure correspondence.

- (12) a. {ACC,3,SG,M} → /lə/, preceding *voit*  
       b. {ACC,3,SG,M} → (↑ OBJ PRED)='PRO'  
                                   (↑ OBJ CASE)=ACC  
                                   (↑ OBJ NUM)=SG  
                                   (↑ OBJ PERS)=3  
                                   (↑ OBJ GEND)=M

So, in this view, the word *le voit* is lexically assigned the syntactic features of *voit*, as well as the syntactic features in (12b). Since this word carries the f-structure information of the object, the use of this word satisfies the object requirement of the verb and precludes the appearance of an NP object for the same reasons noted for the clitic-as-word approach.

### 2.2.2 Agreement vs. pronoun; clitic doubling

In many cases, a clitic is not in perfect complementary distribution with the corresponding phrasal expression, but some amount of clitic doubling is found. In European Spanish, clitic doubling with direct objects (or accusative objects) is found only with pronominal expressions: a definite pronominal direct object is obligatorily expressed as a clitic, optionally doubled by the phrasal expression, as in (13):

- (13) European Spanish (Andrews 1990: 540)

*Alex Alsina*

- a. Lo                vimos            (a él).  
                      3SG.M.ACC see.PST.1PL A HIM  
                      ‘We saw him/HIM.’
- b. \*vimos a él.

In contrast, Rioplatense Spanish (also called Porteño and River Plate Spanish), as well as other varieties of South American Spanish, has a much more general use of direct object clitic doubling, as in (14):

(14) Rioplatense Spanish (Estigarribia 2005)

- a. Yo las                tenía                guardadas las cartas.  
                      I 3PL.F.ACC have.PST.1SG stored            the letters  
                      ‘I had the letters stored.’
- b. ¿La                vas            a llamar a Marta?  
                      3SG.F.ACC go.2SG A call            A Marta  
                      Are you going to call Marta?

In these cases, there is a single GF corresponding to the direct object, which is encoded in the c-structure by both the direct object clitic and by its phrasal expression: *las* and *las cartas* in (14a) and *la* and *a Marta* in (14b). The standard way of analyzing clitic doubling is to assume that it is a kind of agreement: the clitic in examples such as (14) merely specifies the formal features of person, number and gender of the object, while the corresponding phrasal expression contributes, in addition, the semantic PRED feature of the object. This means that there are two sets of specifications associated with clitics that have the dual function exemplified in (13a): as the sole expression of the object, the clitic is lexically associated with the [PRED ‘PRO’] feature needed to satisfy Completeness; as an agreement marker, the clitic lacks this feature in its set of lexical specifications, enabling it to satisfy Uniqueness. This choice is assumed regardless of whether clitics are treated as words, as morphemes, or as exponents.

This analysis follows the treatment given in Bresnan & Mchombo (1987) to subject markers (SM) in Chicheŵa, in contrast with object markers (OM). (See also Fassi Fehri 1984, 1988.) OMs in Chicheŵa are assumed to be always incorporated object pronouns and thus are lexically associated with the [PRED ‘PRO’] feature. SMs in Chicheŵa are claimed to be alternatively pronouns and agreement markers, which follows from the optional [PRED ‘PRO’] feature in the sublexical entry of the SM. Andrews (1990) adapts this idea to the analysis of clitic doubling in Spanish. This way of analyzing the dual function of clitics is used in Mayer (2006)



for Limeño Spanish, in Estigarribia (2013) for Rioplatense Spanish, and in Barbu & Toivonen (2018) for Romanian, among others.<sup>7</sup>

The two sets of lexical specifications associated with clitics that have the dual function just mentioned may differ in more features that in the presence or absence of the [PRED 'PRO'] feature. In Estigarribia (2013), it is proposed that the doubling use of the direct object clitic in Rioplatense Spanish not only lacks the pronominal feature, but carries a constraint that the object cannot be non-specific. The pronominal (or non-doubling) use of the clitic is necessarily definite and specific, but a direct object clitic can double (or agree with) an NP with specific reference (not necessarily definite).

Andrews (1990) explains the facts of European Spanish illustrated in (13) by assuming that direct object clitics also have two lexical entries: the pronominal entry, with the [PRED 'PRO'] feature, and the doubling entry, which has a constraining [PRED 'PRO'] specification, instead of the defining one. This constraining specification effectively restricts the doubling use to situations in which the clitic doubles a pronominal phrase (such as *a él*, in (13a)). The obligatoriness of the clitic double with pronominal NPs is explained by appealing to Andrews's Morphological Blocking Principle. Without this principle, the clitic double would just be an option with pronominal object NPs.

Given two lexical items L1 and L2 such that L1's f-structure specifications are a proper subset of those of L2, the Morphological Blocking Principle requires the use of L2 – the more highly specified lexical item – in a structure in which both L1 and L2 are compatible. In order for this principle to be able to choose between a verb form with a clitic and the same verb form without that clitic, it is necessary to assume that a verb form with a clitic is a lexical item. In other words, the Morphological Blocking Principle presupposes the affixal status of clitics. Given that a clitic is always associated with a set of f-structure features not present in the verb form to which it attaches, a lexical item consisting of a verb and a clitic is always going to be more highly specified in terms of f-structure features than the same lexical item without the clitic. So, if the lexical item with the clitic can be used, it must be used. This explains the obligatoriness of the clitic double in cases like (13).

Constructions that are similar to clitic doubling but that need to be distinguished from it are clitic left dislocation and clitic right dislocation. Languages

---

<sup>7</sup>Although this section deals with object clitics, it should be mentioned that subject clitics, in those languages that have them (e.g. French and Northern Italian languages), also vary as to whether they function as pronouns or as agreement markers, depending on the language and on the context (Poletto & Tortora 2016). See Cardinaletti & Repetti (2010) for the claim that subject clitics in Northern Italian languages should be analyzed as pronouns.

*Alex Alsina*

that do not have clitic doubling or make a very restricted use of it allow these dislocation constructions quite freely. Catalan, which only allows clitic doubling of the direct object in pronominal cases, does not allow a doubling clitic with a neutral intonation in an example like (15a), but allows a direct object clitic, and in fact requires it, when the apparent direct object is fronted, as in (15b), or post-posed, with a clear intonational break, as in (15c):<sup>8</sup>

(15) Catalan (Vallduví 2002: 1233–1237)

- a. (\*El)        va        regalar el llibre a la biblioteca.  
3SG.M.ACC PST.3SG give.INF the book A the library  
‘She/he gave the book to the library.’
- b. El llibre, el        va        regalar a la biblioteca.  
the book 3SG.M.ACC PST.3SG give.INF A the library  
‘The book, she/he gave it to the library.’
- c. El        va        regalar a la biblioteca, el llibre.  
3SG.M.ACC PST.3SG give.INF A the library the book  
‘She/he gave it to the library, the book.’

The left or right dislocations in (15) fulfill functions at the information-structure level, but from the f-structure point of view the dislocated phrase does not fill an in-clause GF, but should be analyzed as a UDF (unbounded dependency function). In other words, the phrase *el llibre* ‘the book’ is not an object in either (15b) or (15c), but a UDF anaphorically bound to the object clitic *el*. It is this element that fulfills the accusative object function in these examples.

### 2.2.3 Non-argument clitics

While clitics in most cases either fulfill a GF that is an argument of the clause or agree with it, there are many instances in which clitics are neither an argument nor a marker of agreement with an argument. This is the case with the reflexive clitic, which can have an inherent use (see Section 2.2.3.1), an anaphoric use

<sup>8</sup>The periphrastic past perfect tense in Catalan consists of an auxiliary form, such as *va* in (15), and an infinitive. The auxiliary is diachronically descended from the present indicative tense of *anar* ‘go’, but synchronically it is not the same form. The past tense auxiliary has the forms *vaig* or *vàreig* (1SG), *vas* or *vares* (2SG), *va* (3SG), *vam* or *vàrem* (1PL), *vau* or *vàreu* (2PL), and *van* or *varen* (3PL), whereas the present indicative of *anar* ‘go’ has the forms *vaig* (1SG), *vas* (2SG), *va* (3SG), *anem* (1PL), *aneu* (2PL), and *van* (3PL). For this reason, the past tense auxiliary is not glossed as if it were a form of *anar* ‘go’.

(Section 2.3), and a use as a marker of passivization or impersonalization (Section 3.2). We shall focus here on two non-argument uses of clitics, leaving aside the reflexive clitic: (a) inherent clitics; and (b) clitics as adjuncts.<sup>9</sup>

### 2.2.3.1 Inherent clitics

Inherent clitics cannot alternate with a phrasal expression and their semantic contribution is not compositional: the predicate consists of a verb and a specific clitic or clitic combination. Examples of verbs with inherent clitics in Catalan include the following: *dinyar-la* ‘die’, *tocar-hi* ‘have a grasp of things’, *anar-se’n* ‘go away’, *jugar-se-la* ‘take a risk’, etc. Without the clitic or clitics, the verb either does not exist (e.g. *dinyar*) or has a different meaning and argument structure (e.g. *tocar* ‘touch’). While one might like to think of these clitics as affixes attached to their verb, they cannot be treated as inseparable affixes, since they can appear separated from the verb by a number of auxiliaries and restructuring verbs, as in the following examples, where the verb and its associated clitics are underlined:

#### (16) Catalan

- a. L’            hauries            poguda            dinyar.  
 3SG.F.ACC have.COND.2SG could.PTCP.F.SG die.INF  
 ‘You could have died.’
- b. Se l’            està    començant a jugar.  
 REFL 3SG.F.ACC be.3SG beginning to play.INF  
 ‘He is beginning to take a risk.’

These examples show that the word to which the inherent clitics attach is not the verb that must be used in combination with these clitics. The string of auxiliaries and restructuring verbs in (16) is clearly not a word, but a sequence of verbs, each one imposing a form requirement on the next. For example, the auxiliary *haver*, in the form *hauries* in (16a), requires the following verb to be in the past participle form, and the verb *poder*, in the form *poguda*, requires the following verb to be in the infinitive form. In addition, *poguda* is in the feminine singular

<sup>9</sup>In addition, one can argue that the clitic *en/ne* found in Catalan, French, and Italian has two other non-argument uses: the partitive use and the genitive use. In the partitive use, the clitic appears instead of the head noun of an object of the verb (see Alsina & Yang 2018 for an analysis of the partitive clitic in Catalan) and cannot be argued to substitute for the whole object. In the genitive use, it fills the complement of a nominal or adjectival complement of the verb and therefore does not correspond to an argument of the verb. Because of space limitations, I will not discuss these uses further.

Alex Alsina

form (as opposed to the unmarked *pogut*) showing agreement with the feminine singular clitic *la*, which in this respect behaves like a direct object. The position in which inherent clitics are realized and the possibility of triggering past participle agreement, among other facts, are the same as with any other clitic.

One might assume that verbs with inherent clitics are listed in the lexicon with one or more fully specified GFs that have no semantic content. For example, *dinyar-la* would fully specify an accusative object with no correspondence to an argument at a-structure or to a semantic participant.<sup>10</sup> It would be listed as the verb *dinyar* taking a feminine singular accusative object, as indicated in (17):

(17) *dinyar*      V      
$$\left[ \begin{array}{c} \text{PRED} \text{ 'DIE<ARG>'} \\ \text{OBJ} \left[ \begin{array}{cc} \text{CASE} & \text{ACC} \\ \text{PRED} & \text{'PRO'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \\ \text{GEND} & \text{FEM} \end{array} \right] \end{array} \right]$$

Under a realizational approach to clitic morphology, we can assume that the features of the object in (17) are mapped onto the clitic *la*. As for the position of this clitic in a string of restructuring verbs, it is no different from that of any clitic. The syntactic dependents of the most embedded verb following a string of restructuring verbs can cliticize onto the highest verb in the string of verbs.

One of the uses of the reflexive clitic is as an inherent clitic. In this use, there is no reflexive interpretation. Following Grimshaw (1982), we can distinguish two classes of verbs within the class of verbs that take an inherent reflexive clitic: the lexically stipulated class of reflexive verbs and the class of inchoative verbs (in Grimshaw’s terminology). The first class consists of verbs that are lexically required to take a reflexive clitic and either do not exist in a non-reflexive form or are not related in a systematic way with their non-reflexive counterpart. Examples of this class in Catalan are *desmaiar-se* ‘faint’ or *penedir-se* ‘repent’, which do not exist without a reflexive clitic. In the second class we find the intransitive alternant of the causative alternation, such as *trencar-se* ‘break.INTR’ or *obrir-se* ‘open.INTR’ in Catalan. See Alsina (2020) for a treatment of inherently reflexive verbs.

<sup>10</sup>One could debate whether this object should have a [PRED ‘PRO’] feature. Depending on how one views the syntax-morphology mapping for clitics, this feature might be necessary. On the other hand, the presence of this feature on a non-semantic GF would yield a violation of Coherence, according to some definitions of this condition which require a PRED feature on all and only those GFs with semantic content.

## 2.2.3.2 Clitics as adjuncts

Although clitics generally correspond to objects (or subjects, in languages with subject clitics, such as French), in some languages they can also correspond to obliques: this is the case of *en* and *y* in French, *en* and *hi* in Catalan, and *ne* and *ci* or *vi* Italian. The clitic *y/hi/ci(vi)* may correspond either to an argument of the verb or to an adjunct, as we see in (18) for French and in (19) for Catalan:

## (18) French (Schwarze 2001b)

- a. J' y ai            pensé.  
I y have.1SG thought  
'I have thought of it.'
- b. Je l'            y ai    vu.  
I 3SG.F.ACC Y have seen  
'I saw him there.'

## (19) Catalan (Todolí 2002)

- a. Encara no s'    hi han    acostumat.  
still    not REFL HI have.3PL accustomed  
'They haven't got used to it yet.'
- b. No es    pot circular sense    casc,    però molts motoristes    hi  
Not REFL can ride.INF without helmet, but    many motorcyclists HI  
circulen.  
ride.3PL  
'You cannot ride without a helmet, but many motorcyclists do so.'

In (18a) and (19a), *y/hi* corresponds to an argument, but in the b examples it is an adjunct: in (18b) it expresses the location in which an event takes place, and in (19b) it expresses the means or manner. One can take this to mean that *y/hi* has a double function, being alternatively an oblique or an adjunct, as in Schwarze (2001b). Or one can take this as evidence that there is no adjunct grammatical function, as argued in Alsina (1996b). According to Alsina (1996b), the distinction between argument and adjunct is made at the level of a-structure: a GF that corresponds to a position at the a-structure is an argument, whereas a GF with semantic content that does not is an adjunct. This distinction need not be duplicated at the level of GFs by increasing the inventory of GFs with ADJ, and adjuncts are simply obliques (OBL) at the level of GFs. Consequently, all we need to say about *hi/y* is that it corresponds to an OBL. By not restricting it to arguments, it follows that it can correspond to either an argument or an adjunct.

Alex Alsina

## 2.3 The anaphoric reflexive clitic

We can define reflexive clitics as those that show agreement in person and number with the logical subject<sup>11</sup> of the predicate that the clitic combines with. First and second person clitics do not have a special reflexive form distinct from their non-reflexive form. The third person does have a specific form for the reflexive use, *se* (and cognate forms), which, however, does not distinguish singular from plural. The third person form, being the only one that is unambiguously reflexive, will be normally used to illustrate the behavior of reflexive clitics.

In this section, we will only consider what we might call the anaphoric use of the reflexive clitic, by which the predicate has a semantically reflexive or reciprocal interpretation. In Section 2.3.1, we compare the pronominal analysis and the valence-reducing analysis of the anaphoric reflexive. And in Section 2.3.2, we consider three variants of the valence-reducing analysis.

The other uses of the reflexive clitic are the inherent use (Section 2.2.3) and the passive and impersonal use (Section 3.2).<sup>12</sup>

### 2.3.1 The reflexive clitic as an argument or as a marker of valence-reduction

In general, any verb that can take an object (direct or indirect) can also take a reflexive clitic instead of the phrasal object, so that the logical subject and another direct argument of the verb are interpreted as being the same set of participants: This is the anaphoric use of the reflexive clitic. The interpretation is reflexive or reciprocal depending on whether the same participant (individual or group) is involved in the relation – reflexive interpretation – or a different participant of the set is involved – reciprocal. Using Catalan to exemplify the anaphoric use of the reflexive clitic, (20a) is a transitive sentence in which the direct, or accusative, object is expressed as an NP; (20b) shows that a reflexive clitic can be used instead of the NP object, in this case with a reflexive interpretation; and this sentence resembles (20c), where a pronominal non-reflexive clitic is used instead of the object NP. The examples in (21) show the possibility of the reflexive

<sup>11</sup>See the glossary for the definition of *logical subject*.

<sup>12</sup>The homonymy or syncretism of the anaphoric reflexive with the passive/impersonal reflexive is complete in some Romance languages (e.g. Spanish, Catalan, or French), but is not complete in some others, specifically, in Italian. In Italian, in both uses, it has the form *si* when it is not in combination with another clitic, but, when the two uses co-occur in the same clause, we obtain the combination *ci si*, as in *ci si lava* in (4c). In addition, the anaphoric reflexive precedes a third person accusative clitic, whereas the impersonal reflexive follows it, as shown in (4). This indicates that they are different morphs in Italian, which explains the possibility of their co-occurrence together with another clitic, as *ce lo si compra* ‘one buys it for oneself’, as pointed out by an anonymous reviewer.

clitic appearing instead of a dative object and yielding a reciprocal or reflexive interpretation.

(20) Catalan

- a. Mira com contradiu el director.  
look how contradict.3SG the manager  
'See how she contradicts the manager.'
- b. Mira com es contradiu.  
look how REFL contradict.3SG  
'See how she contradicts herself.'
- c. Mira com el contradiu.  
look how 3SG.M.ACC contradict.3SG  
'See how she contradicts him.'

(21) Catalan

- a. Avui els estudiants enviaran regals a la professora.  
today the students send.FUT.3PL presents A the teacher  
'Today the students will send the teacher presents.'
- b. Avui els estudiants s' enviaran regals.  
today the students REFL send.FUT.3PL presents  
'Today the students will send each other/themselves presents.'
- c. Avui els estudiants li enviaran regals.  
today the students 3SG.DAT send.FUT.3PL presents  
'Today the students will send her presents.'

This pattern of facts lends itself to an analysis in which the reflexive clitic only differs from pronominal object clitics in its anaphoric properties, being obligatorily bound by some antecedent in a local domain, and is the realization of an argument of the clause. This is in fact the analysis proposed in Alencar & Kelling (2005), which we can call the "pronominal analysis." In examples like (20b) and (21b), the reflexive clitic would be argued to realize an accusative object or a dative object, just like the non-reflexive clitics do. However, this analysis has been shown to be problematic since Grimshaw (1982). Grimshaw (1982, 1990) gives compelling evidence for the claim that the reflexive clitic in its anaphoric use should be treated as not realizing an argument of the clause but as valence-reducing morphology.

The clearest evidence presented by Grimshaw (1982, 1990) for the valence-reducing analysis of the reflexive clitic concerns the behavior of the causative

*Alex Alsina*

construction. The logical subject of the infinitive in a causative construction, with *faire* in French, is realized differently depending on the transitivity of the infinitive: indirect object if the infinitive has a direct object, and direct object otherwise, as shown in (22):

(22) French (Grimshaw 1990: 153)

- a. Il fera boire un peu de vin \*(à) son enfant.  
 he make.FUT.3SG drink.INF a bit of wine A his child  
 ‘He will make his child drink a little wine.’
- b. Il fera partir {les/\*aux} enfants.  
 he make.FUT.3SG leave.INF the/\*A.the children  
 ‘He will make the children leave.’

When the infinitive has a reflexive clitic corresponding to its direct object, it behaves like an intransitive verb and its logical subject is realized as a direct object, as in (23a). In contrast, if the direct object of the infinitive is expressed as a non-reflexive clitic, its logical subject is an indirect object, as in (23b).

(23) French (Grimshaw 1990: 153)

- a. La crainte du scandale a fait se tuer {le/\*au} frère  
 the fear of.the scandal has made REFL kill.INF the/\*A.the brother  
 du juge.  
 of.the judge  
 ‘Fear of scandal made the brother of the judge kill himself.’
- b. La crainte du scandale l’a fait tuer {au/\*le}  
 the fear of.the scandal 3SG.M.ACC.has made kill.INF A.the/\*the  
 juge.  
 judge  
 ‘Fear of scandal made the judge kill him.’

If we assume that the reflexive clitic is not an object, unlike the non-reflexive clitic, but an element of the morphology that signals the binding of two arguments so that there is only one open argument position, we explain that the verb behaves like an intransitive verb.

Grimshaw (1982) also presents NP extraposition in French as evidence for the intransitive behavior of reflexivized verbs, i.e., verbs with an anaphoric reflexive clitic. French allows arguments that can normally appear as subjects, as in (24a), to alternatively appear as objects with a dummy *il* in subject position, as in (24b):



- (24) French (Grimshaw 1982: 112)
- a. Un train passe toutes les heures.  
a train passes all the hours
  - b. Il passe un train toutes les heures.  
IL passes a train all the hours  
'A train goes by every hour.'

However, the construction of NP extraposition, illustrated in (24b), is restricted to intransitive verbs. In addition, there are semantic constraints on NP extraposition, but the intransitivity requirement is independent of these semantic restrictions. A reflexivized verb behaves like an intransitive verb in allowing NP extraposition, unlike verbs with non-reflexive object clitics, as the contrast in (25) illustrates:

- (25) French (Grimshaw 1982: 113)
- a. Il s' est dénoncé trois mille hommes ce mois-ci.  
IL REFL is denounced three thousand men this month  
'Three thousand men denounced themselves this month.'
  - b. \*Il l' a dénoncée trois mille hommes.  
IL 3SG.F.ACC has denounced three thousand men  
'Three thousand men denounced it.'

### 2.3.2 Three alternative valence-reducing analyses

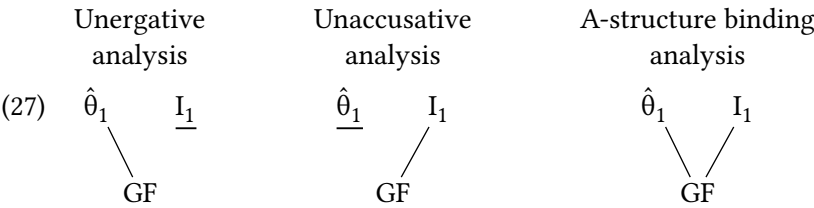
Having shown that reflexive cliticization turns a transitive verb into an intransitive one, three possibilities emerge as to how the two argument roles involved in the binding relation signaled by the reflexive clitic map onto only one GF (typically the subject, but not necessarily, as shown in (25a)). The three analyses, described in (26), have in common the idea that the anaphoric reflexive clitic signals the binding at the level of argument structure of the logical subject and another core argument of the same predicate:

- (26) a. *The unergative analysis*: the lower argument is lexically bound and therefore unable to be expressed as a GF; only the logical subject is expressed as a GF. Proposed by Grimshaw (1982).
- b. *The unaccusative analysis*: the logical subject is lexically bound and therefore unable to be expressed as a GF; only the lower argument in the binding relation is expressed as a GF. Proposed by Grimshaw (1990).

Alex Alsina

- c. *The a-structure binding analysis*: both arguments involved in the binding relation are expressed as a GF and are expressed as the same GF. Proposed by Alsina (1993, 1996b).

Schematically, the three analyses can be depicted as in (27), where “ $\hat{\theta}$ ” represents logical subject, “I” represents internal argument, co-subscripting signifies binding of arguments, and underlining of an argument signifies that the argument has no mapping to GF:



Grimshaw (1982) does not present evidence specifically for the unergative analysis. The evidence presented in Grimshaw (1990) for the unaccusative analysis rests primarily on the facts of auxiliary selection in Italian, as we shall see. Some of the evidence presented in favor of this analysis is really neutral with respect to the other two analyses in competition. Since, according to Grimshaw (1990: 154), reflexivization satisfies an external argument (by binding), it cannot apply to predicates that do not have an external argument or have a suppressed external argument. It follows that it cannot apply to passives or subject-raising predicates. This explains the contrast between English and French with subject-raising verbs (from Grimshaw 1990: 155):

- (28) a. They appear to each other to be intelligent.  
      b. \*Jean se     semble intelligent. (French)  
          Jean REFL seems intelligent.  
          ‘Jean seems intelligent to himself.’

Grimshaw (1990) takes the ungrammaticality of (28b) to follow from the assumption that a raising verb like *sembler* ‘seem’ does not have an external argument. However, it can also be attributed to the observation that this verb does not have two arguments that can be involved in binding: the subject in (28b) is not an argument of the raising verb, but of its complement, so that the two arguments that would be involved in binding in (28b) belong to two different predicates. And the three analyses described in (26)–(27) require that the two arguments involved in reflexive cliticization be arguments of the same predicate.

As for auxiliary selection in Italian, unergative verbs select *avere* ‘have’ as the auxiliary in perfective compound forms and unaccusative verbs select *essere* ‘be’ (following Perlmutter 1978, 1983, 1989 and Rosen 1984; see Loporcaro 2016 for an update), as shown in (29). The fact that reflexivized verbs select *essere*, as in (30), even though their non-reflexive counterparts select *avere*, is taken as evidence in Grimshaw (1990) that reflexivized verbs are unaccusatives:

(29) Italian (Katerinov 1975)

- a. Avete viaggiato bene?  
have.2PL travelled well  
‘Have you travelled well?’
- b. Sono uscito.  
be.1SG gone.out  
‘I have gone out.’

(30) Italian (Katerinov 1975)

- a. Maria e Paola si sono salutate.  
Maria and Paola REFL be.3PL greeted.F.PL  
‘Maria and Paola greeted each other.’
- b. Mi sono comprato una casa nuova.  
1SG be.1SG bought a house new  
‘I bought myself a new house.’

If the expressed argument in reflexivized verbs is the internal argument, and the external argument is not assigned to a GF, as in the unaccusative analysis in (27), it is clear that reflexivized verbs are like unaccusative verbs. However, let us suppose that the relevant notion for auxiliary selection is that verbs whose highest GF maps onto an internal argument select *essere* (where SUBJ ranks higher than OBJ, and OBJ than OBL). Then, both the unaccusative analysis and the a-structure binding analysis fare equally in predicting that both unaccusative verbs and reflexivized verbs select *essere*.

But the a-structure binding analysis does not treat reflexivized verbs as unaccusatives, since the highest GF of the former is an external argument, as well as an internal argument. This has an advantage over the unaccusative analysis as it allows us to explain two facts that the unaccusative analysis fails to explain. First, the highest GF of reflexivized verbs, being an external argument, tends to be a subject much more so than that of unaccusative verbs, which is not an external argument. This contrast between reflexivized verbs and unaccusative verbs can

Alex Alsina

be clearly illustrated by using the same verb with a reflexive clitic yielding a potential ambiguity between the anaphoric and the passive interpretations. Using Catalan data, a sentence like (31a) is ambiguous between these two interpretations, whereas (31b) only allows the anaphoric interpretation:

(31) Catalan

- a. Es defensaran dos diputats al parlament.  
REFL defend.FUT.3PL two deputies at.the parliament  
'Two deputies will defend themselves at the parliament.'  
'Two deputies will be defended at the parliament.'
- b. Dos diputats es defensaran al parlament.  
two deputies REFL defend.FUT.3PL at.the parliament  
'Two deputies will defend themselves at the parliament.'  
\* 'Two deputies will be defended at the parliament.'

The preverbal position of the NP, with no object clitic anaphorically dependent on it attached to the verb, unambiguously signals that the NP is the subject – or, more exactly, a topic anaphorically linked to the null pronominal subject. While an internal argument, especially if expressed as an indefinite NP, is assigned the object function, an external argument favors the assignment to the subject function.

The contrast between the reflexivized verb and the reflexive passive form is even clearer, when, under the appropriate discourse conditions, we omit the noun *diputats* from (31). If the NP *dos* is postverbal, with obligatory presence of the partitive clitic *en*, only the passive interpretation is allowed; if the NP *dos* is preverbal, with no partitive clitic, only the reflexivized reading is possible:

(32) Catalan

- a. Se' n defensaran dos al parlament.  
REFL EN defend.FUT.3PL two at.the parliament  
\* 'Two will defend themselves at the parliament.'  
'Two will be defended at the parliament.'
- b. Dos es defensaran al parlament.  
two REFL defend.FUT.3PL at.the parliament  
'Two will defend themselves at the parliament.'  
\* 'Two will be defended at the parliament.'

If, as assumed in Grimshaw (1990), the reflexive passive and the reflexivized verb have the same syntactically expressed arguments, namely, the internal argument

in both cases, the difference shown in (31) and (32) would be completely unexpected. On the other hand, under the a-structure binding analysis of reflexivized forms, these forms have a GF that is both an internal and an external argument, contrasting with reflexive passive forms, in which the highest GF is only an internal argument.

The second fact that favors the a-structure binding analysis is found in triadic predicates: When the binding relation involves an argument that in the non-reflexivized form of the verb is a dative object, the corresponding GF is not dative in the reflexivized form, but nominative. If argument realization with reflexivized verbs were the same as with unaccusative or passive verbs, we would not expect dative case to disappear. Dative case is retained under passivization, blocking the dative expression from being the passive subject. We see this not only with participial passives, but also with reflexive passives, as in (33b). The goal argument is dative and cannot be expressed as a nominative phrase in a reflexive passive, as in (33c). However, in the reflexivized form, in (33d) with a reciprocal interpretation, the goal argument is nominative and the subject.<sup>13</sup>

(33) Catalan

- a. El metge va ensenyar els resultats al pacient.  
the doctor PST.3SG show.INF the results A.the patient  
'The doctor showed the patient the results.'
- b. Es van ensenyar els resultats al pacient.  
REFL PST.3PL show.INF the results A.the patient  
'The patient was shown the results.'
- c. \*El pacient es va ensenyar els resultats.  
the patient REFL PST.3SG show.INF the results  
'The patient was shown the results.'
- d. Els pacients es van ensenyar les cicatrius.  
the patients REFL PST.3PL show.INF the scars  
'The patients showed each other the scars.'

Under the unaccusative analysis, the NP *els pacients* in (33d) is the goal internal argument, just as the phrase *al pacient* in (33b); so, it is very unclear why it has dative case in the passive example, which prevents it from being the subject, as

<sup>13</sup>The phenomenon is illustrated with Catalan data, but the facts are essentially the same in French, Italian, and Spanish. See, for example, the Italian reflexivized form (30b), where the first person singular reflexive clitic signals the binding of the agent and the goal, which are encoded as the (null) subject.

Alex Alsina

in (33c), but not in the reflexivized form, in which the goal argument is nominative.<sup>14</sup> On the other hand, within the a-structure binding analysis, the phrase *al pacient* in the passive example (33b) is the goal internal argument and no other argument, whereas the phrase *els pacients* in the reflexivized structure (33d) is both the goal internal argument and the external argument. Here there are two arguments that map onto the same GF. If we assume, as in Alsina (1996b), that dative case is assigned to the GF that maps onto the more prominent of two internal arguments, as long as it is **not an external argument**, it follows that dative case will be assigned to the goal internal argument in the active and passive forms (33a) and (33b), but not in the reflexivized form (33d).

The a-structure binding analysis of the anaphoric use of the reflexive clitic just described relies on the idea essential to LFG that grammatical information is factored into different levels of representation, allowing for mismatches among these levels. In particular, the distinction between argument roles at a-structure and GFs at f-structure plays a crucial role in this analysis. If we allow for the possibility that a given GF corresponds to two different argument roles, as schematized in (27) for the a-structure binding analysis, we can explain not only the valence-reducing effect of the anaphoric reflexive clitic, but those properties of the GF that group it with an internal argument, as in the unaccusative analysis, and those properties that group it with an external argument, as in the unergative analysis.

Following the proposal in Alsina (1996b), we can illustrate this by comparing the non-reflexive use of a dyadic predicate such as *defensar* ‘defend’ in Catalan with the same predicate with the anaphoric reflexive clitic. This predicate has an external argument and an internal argument, represented by [Ext] and [Int] respectively at a-structure. Each argument has its linking index, represented as a subscripted number, which, in the default case, is different for each argument, entailing a different mapping to GF. This is the situation in (34a), where the external argument maps onto the subject and the internal argument onto the object. The effect of the anaphoric reflexive clitic is to coindex the logical subject of a predicate with an internal argument, so that they have the same linking index and therefore map onto the same GF, as shown in (34b). The principles mapping argument roles to GFs are satisfied in (34b): the external argument is required to map onto the subject and the internal argument is required to map onto a direct GF (either subject or object) and, since the subject is a direct GF, both mapping requirements are met. The a-structure is represented as the value of the feature PRED in (34).

<sup>14</sup>Grimshaw (1990: 184) points out this problem in an endnote and essentially leaves it unsolved, although one of the solutions she sketches involves precisely a-structure binding.

- (34) a. Non-reflexive use of *defensar* ‘defend’:  

$$\left[ \begin{array}{l} \text{PRED 'DEFEND'} \langle [\text{EXT}]_1 [\text{INT}]_2 \rangle' \\ \text{SUBJ}_1 \\ \text{OBJ}_2 \end{array} \right]$$
- b. Reflexivized use of *defensar-se* ‘defend-REFL’:  

$$\left[ \begin{array}{l} \text{PRED 'DEFEND'} \langle [\text{EXT}]_1 [\text{INT}]_1 \rangle' \\ \text{SUBJ}_1 \end{array} \right]$$

### 3 Arguments, grammatical functions, and case

This section deals with the morphosyntactic expression of arguments in terms of grammatical functions and case. Section 3.1 considers the inventory of GFs, especially the GFs of subjects, objects, and clausal complements. The passive and impersonal reflexive constructions are examined in Section 3.2.

#### 3.1 Objects and their realization

##### 3.1.1 Direct and indirect objects: GF and case

Traditional grammar, as well as Relational Grammar, distinguishes two kinds of objects in the Romance languages: direct object (DO) and indirect object (IO). DOs, in their phrasal expression, are generally NPs without any case marker or preposition, except that in some languages a subset of DOs are marked by a preposition,<sup>15</sup> whereas IOs, as phrases, are PPs introduced by the preposition *a*. Both kinds of objects can be expressed as clitics and all Romance languages have different sets of pronominal clitics in the third person for the two kinds of objects. First and second person clitics do not distinguish between the two kinds of objects.<sup>16</sup> Given that LFG does not have a DO and an IO in its standard inventory of GFs, researchers have accommodated this distinction into the LFG inventory of GFs in different ways. The proposals that restrict themselves to the standard LFG inventory of GFs have in common the assumption that the DO is

<sup>15</sup>The prepositional marking of the DO, also known as differential object marking, is found in Spanish, Catalan, southern Italian dialects, and Sardinian, which use the same preposition as for IOs, and in Romanian, in which the preposition *pe* is used (Dragomirescu & Nicolae 2016: 920–921). See Barbu & Toivonen (2018) for the distribution of DO *pe* in Romanian.

<sup>16</sup>Neither do third person reflexive clitics, but then, according to Section 2.3, they are not object clitics. Instances of DO-IO syncretism are found even in third person non-reflexive clitics: this is the case of Spanish *leísmo*, in which the clitic *le* is used for both IOs and human masculine DOs. Other forms of DO-IO syncretisms in third person clitics are found in regional varieties of Spanish (Tuten et al. 2016: 398).

*Alex Alsina*

OBJ and differ in the GF attributed to the IO, which is one of the following three: OBL, OBJ<sub>θ</sub>, and OBJ.<sup>17</sup>

### 3.1.1.1 IO as OBL

This proposal is found in Schwarze (2001b) and Sells (2013: 185–194), although no motivation is given for adopting it instead of the available alternatives. Alsina (1996b: 150–160) enumerates eight properties that group IOs with DOs, in the class of direct functions, together with subjects, contrasting them with obliques: (1) doubling of independent personal pronouns in the verbal morphology (as clitics); (2) expression of person and number distinctions in the verbal morphology; (3) the ability to be bound at a-structure (by means of the reflexive clitic); (4) the ability to launch a floating quantifier; (5) disjoint reference of pronouns; (6) the ability to bind quantifiers; (7) the ability of independent (or strong) pronouns to function as resumptive pronouns; and (8) the ability to be the target of secondary predication. All of these properties argue against treating the IO as an oblique and show that it belongs to the class of direct GFs, together with subjects and objects.<sup>18</sup>

### 3.1.1.2 IO as OBJ<sub>θ</sub>

This proposal is found in Falk (2001: 115–118), Alencar & Kelling (2005), Aronovich (2012), Quaglia (2012), and Carretero García (2018). Grimshaw (1982) can be grouped in this proposal, as she assumes that the DO is OBJ and uses the GF A OBJ, instead of OBJ<sub>θ</sub>, for the IO. The main argument for this proposal is the observation that dative arguments cannot be encoded as subjects: they are never the subject of a passive form, with verbs that can be passivized, and are not the subject of psychological verbs of the ‘like’-type. While this is true, there are many reasons for rejecting this proposal. In languages such as Chicheŵa (asymmetrical object languages), in which the OBJ-OBJ<sub>θ</sub> distinction is strongly motivated, the OBJ has the ability to be expressed as a morphologically incorporated pronominal, can be accessed by an a-structure binding operation (reciprocalization), and

<sup>17</sup>Some exceptions to this observation are found. Luís & Otoguro (2004: 344–349) treat the single object of a clause as OBJ, whether it is direct or indirect (i.e., accusative or dative) and, in ditransitive clauses, treat the DO as OBJ<sub>θ</sub> and the IO as OBJ. Luís & Spencer (2005) use the GFs OBJ1 and OBJ2 for the IO and the DO respectively, where we can assume that OBJ1 is another name for OBJ and OBJ2 replaces OBJ<sub>θ</sub>. No argumentation is presented for these proposals.

<sup>18</sup>To these properties we could add the IO-DO syncretism in first and second person and reflexive clitics in Romance in general, the partial IO-DO syncretism in third person non-reflexive clitics in Spanish (see footnote 16), and the partial syncretism in the phrasal expression of IO and DO in those languages that use the same preposition for both objects (see footnote 15).



alternates with the SUBJ in a passive form, whereas the OBJ<sub>θ</sub> lacks all of these properties (see Baker 1988a,b, Alsina & Mchombo 1990, Bresnan & Moshi 1990, Alsina 1996a, among others).<sup>19</sup> The IO, like the DO, in Romance is able to be expressed as a morphologically incorporated pronoun, as illustrated in examples (3b), (3c) and (21c) (see also (35)), and, like the DO, can be accessed by an a-structure binding operation (by means of the reflexive clitic), as in (21b), (30b), and (33d). The only property that the IO shares with the OBJ<sub>θ</sub> is the fact that it cannot be a subject. To focus on this one feature of the IO in order to claim that it is an OBJ<sub>θ</sub> is to ignore the fact that there is a cluster of properties associated with the OBJ-OBJ<sub>θ</sub> distinction, as has been mentioned, and the fact that DO and IO are distinguished by grammatical case, unlike OBJ and OBJ<sub>θ</sub> in most asymmetrical languages.

In addition to this, there is a difference in the thematic roles that map onto OBJ<sub>θ</sub> in the subclass of asymmetrical languages of the Chicheŵa type termed non-alternating in Alsina (1996a) and the thematic roles that correspond to IO in the Romance languages. In Chicheŵa, only thematic roles below goal in the thematic hierarchy (i.e., instrumental, theme, patient, locative) can map onto OBJ<sub>θ</sub>, as the higher roles in the hierarchy (agent, beneficiary, goal) cannot map to OBJ<sub>θ</sub>. In contrast with this, the IO in Romance typically corresponds to the higher roles in the hierarchy (agent, beneficiary, goal, experiencer).

In other words, to assume that IO is OBJ<sub>θ</sub> implies abandoning the idea that there is a cluster of properties associated with OBJ<sub>θ</sub> and assuming that the only necessary and sufficient condition for the OBJ<sub>θ</sub> function is the failure of alternating with the SUBJ function, which is clearly an undesirable loss of predictive power of the theory. And it also requires assuming that the mapping of argument roles to OBJ<sub>θ</sub> may vary radically from language to language.

### 3.1.1.3 IO as OBJ

This proposal is argued for in Alsina (1996b) and is also found in Vanhoe (2002). It places a lot of importance on the observation that DO and IO are distinguished primarily by means of grammatical case. Both DO and IO are the GF OBJ and are distinguished because IO is dative and DO is non-dative (i.e., accusative, although nominative is also an option, see Section 3.1.3). What needs to be accounted for

<sup>19</sup>In addition, in Chicheŵa, the OBJ precedes the OBJ<sub>θ</sub> when both are expressed as NPs in the VP. However, this is not a necessary property of asymmetrical object languages, as there are languages of this type, including other Bantu languages, that allow either order of the objects. Also, the fact that the DO precedes the IO in Romance in the unmarked order is simply a consequence of the different grammatical category of the two objects, the DO being an NP and the IO being a PP.

*Alex Alsina*

in this approach is case assignment, particularly, the assignment of dative case. Alsina (1996b) notes that dative case is assigned either on the basis of the semantics, specifically, the thematic role involved, or on the basis of the a-structure configuration. In the first case, dative is claimed to be assigned to arguments whose thematic role is goal and this assignment does not depend on there being a non-dative object in the clause, as illustrated in (35a). In the second case, dative is assigned to the GF corresponding to the more prominent of two internal arguments, as in (35b). As there need to be two internal arguments each mapping to a different GF for the latter type of dative case assignment, dative fails to be assigned to the single internal argument of a clause (unless it meets the semantic requirement), as in (35c). The dative-accusative case alternation in (35b)–(35c) also occurs with the causee in causative constructions depending on transitivity of the embedded infinitive (see (22) and Section 4.1).

(35) Catalan (Alsina 1996b: 172)

- a. En Ferran li ha escrit (una carta).  
ART Ferran 3SG.DAT has written a letter  
'Ferran has written him (a letter).'
- b. Li ensenyen llatí.  
3SG.DAT they.teach Latin  
'They teach him Latin.'
- c. L' ensenyen.  
3SG.M.ACC they.teach  
'They teach him.'

The only property that seems to indicate that IO behaves like  $OBJ_{\theta}$  is the claim that dative arguments are never subjects in Romance, but must be objects instead. Alsina (1996b) claims that this fact is best accounted for through a constraint prohibiting subjects with dative case. This constraint is active in the Romance languages, which do not allow dative subjects,<sup>20</sup> but is not active in languages such as Icelandic or Hindi (see Zaenen et al. 1985 and Mohanan 1994, respectively, among others), in which dative subjects are possible. The thematic roles to which dative case is assigned are very similar across these different languages, but Romance differs from Icelandic and Hindi basically because dative blocks the assignment of the subject function in the former, but not in the latter. Introduc-

<sup>20</sup>However, some authors have claimed that dative experiencers can be subjects, e.g. Cardinaletti (2004) for Italian and Fernández-Soriano (1999) for Spanish.

ing the OBJ-OBJ<sub>θ</sub> distinction in the description of the facts would just obscure the differences and similarities among these languages.<sup>21</sup>

Accepting the idea that IO is OBJ implies that a given clause may have more than one GF OBJ, since clauses often have an IO and a DO and sometimes even more than one IO. In this respect, OBJ would not be different from OBJ<sub>θ</sub> or OBL<sub>θ</sub>, of which clauses may have more than one. This requires modifying the framework, which, in its standard form, does not allow multiple GFs with the same attribute, unless the GF in question is assumed to take a set of f-structures as its value rather than a single f-structure. Alsina (1996b) assumes that the only GF that is unique in a clause is the subject, whereas the other two GFs, namely, object and oblique (in a reduced inventory of GFs with only the three named GFs), are not required to be unique and can have multiple instantiations. This proposal can be implemented by assuming that both OBJ and OBL take a set of f-structures as their value, whereas SUBJ takes an f-structure as its value. See also Patejuk & Przepiórkowski (2016) for a different implementation of the idea that the inventory of GFs consists of only the three GFs mentioned.<sup>22</sup>

### 3.1.2 The GF of clausal complements

The debate about the inventory of GFs in LFG has also addressed the issue of the GF COMP, a GF that in standard LFG is reserved for clausal complements, typically finite. Alsina et al. (2005) (AMM) argue that this GF is not necessary and, in fact, complicates the statement of generalizations and that clausal complements should be assumed to be either objects or obliques.<sup>23</sup> The argument based on Catalan is as follows. Catalan has two types of clausal complements introduced by the complementizer *que*, without a preposition: those that alternate with a nominal complement, which can be expressed by the object clitic *ho*, and that can passivize, and those that alternate with a prepositional complement, that can be expressed by one of the oblique clitics *hi* or *en*, and that cannot passivize. (36) exemplifies a complement of the first type: the verb *entendre* ‘understand’

<sup>21</sup>Certain verbs take a dative object as their sole object. This occurs in Latin with verbs such as *subvenire* ‘help’, *parcere* ‘spare’, etc., as well as in the Romance languages. This is unlike the OBJ<sub>θ</sub> in languages such as Chicheŵa, where it occurs only in a double object construction.

<sup>22</sup>This idea is also valid for asymmetrical languages like Chicheŵa, since the distinction between primary and secondary object (OBJ and OBJ<sub>θ</sub>, respectively, in standard LFG) needs to be made at the level of a-structure, as argued in Alsina (1993, 2001), by means of a feature (R) that marks secondary objects, and only at that level, so that both primary and secondary objects are simply objects at the level of GFs.

<sup>23</sup>A defense of the GF COMP can be found in Dalrymple & Lødrup (2000), Lødrup (2004, 2012), and Belyaev et al. (2017).

*Alex Alsina*

can take a nominal complement, as in (36a), can cliticize its clausal complement by means of *ho*, as in (36b), and can passivize with the dependent clause as the subject, as in (36c):

(36) Catalan (AMM)

- a. (La teva explicació) no l'                    he                    entesa.  
the your explanation not 3SG.F.ACC have.1SG understood.F  
'(Your explanation<sub>i</sub>) I didn't understand it<sub>i</sub>.'
- b. (Que hakis    arribat tan tard) no ho he                    entès.  
that have.2SG arrived so late not HO have.1SG understood  
'(That you should have arrived so late<sub>i</sub>) I didn't understand it<sub>i</sub>.'
- c. Que votessis        a favor de la    proposta no va        ser  
that vote.SBJV.2SG in favor of the proposal not PST.3SG be  
entès                per una part del    públic.  
understood by a part of-the audience  
'That you should have voted in favor of the proposal was not  
understood by part of the audience.'

*Convèncer* 'convince' is a verb that takes a clausal complement of the second type: it alternates with a PP, as in (37a), but does not take a preposition, as in (37b), and can be expressed by means of the oblique clitic *en*, as in (37c):

(37) Catalan (AMM)

- a. M' heu        de convèncer **de** les seves possibilitats.  
me have.2PL to convince of the 3POSS possibilities  
'You have to convince me of his possibilities.'
- b. M' heu        de convèncer (\***de**) que torni        a casa.  
me have.2PL to convince of that return.1sg to home  
'You have to convince me to return home.'
- c. Me n' heu        de convèncer.  
me EN have.2PL of convince  
'You have to convince me of that.'

Another class of verbs that take a clausal complement introduced by *que* is illustrated by *estar d'acord* 'agree', which takes a different preposition, *en*, when the complement is not clausal, and a different clitic form, *hi* (see relevant examples in AMM).

The choice of oblique clitic (*en* vs. *hi*) is related to the choice of oblique preposition: oblique complements introduced by the preposition *de* can be expressed by means of the clitic *en*, whereas other obliques alternate with the clitic *hi*. Replacing one oblique clitic by the other one renders the sentences ungrammatical. In addition, neither of the two classes of verbs allows the dependent clause introduced by *que* to be the subject of a passive form, as illustrated in (38) for *convèncer*.

(38) Catalan (AMM)

\*Que tornés                    a casa    va                    ser convençut en Martí.  
that return.SUBJ.3SG to home PAST.3SG be convinced the Martí  
'That he return home was convinced Martí.'

A possible LFG approach to these facts using the COMP function would assume that a clausal complement can be either an OBJ or a COMP: it is an OBJ in cases like (36b), where it alternates with an NP, with object clitics, and with the subject in a passive clause, whereas it is a COMP in (37b), where it has none of these properties. This means that predicates like *convèncer* and *estar d'acord* have two different subcategorization frames depending on whether the complement is nominal or clausal: they take an OBL for sentences such as (37a) and a COMP for sentences such as (37b) and, to complicate matters further, the clitic that corresponds to the OBL and to the COMP is unique for each verb regardless of whether it corresponds to the OBL or to the COMP, as in (37c). No generalization can be made regarding the choice of clitic, given that some COMPS are expressed as *en* and some others are expressed as *hi*, and the choice does not depend on the COMP but on the OBL that appears on the alternative subcategorization frame of the verb.

If, on the other hand, we assume that there is no such GF as COMP, as claimed in AMM, but clauses can be the c-structure realization of either OBJ or OBL (just as they can be of SUBJ), the different behavior of the clausal complements shown in (36)–(38) simply follows from their being either OBJ or OBL, together with a constraint preventing clausal complements from taking a preposition. This constraint (let's call it \*P+CP) is active in languages like Catalan or French (see Forst 2006 for relevant data on French) and English, where clausal complements are not preceded by a preposition, but not in languages like Spanish, where complements take their required preposition regardless of the category of the complement (nominal or clausal).<sup>24</sup>

<sup>24</sup>Danish, according to Nigel Vincent (p.c.), is another language where the \*P+CP constraint is not active: e.g. *det endte med at han blev fyret* 'it ended with that he was fired'.

Alex Alsina

In languages with an active \*P+CP, a verb selecting an oblique with a particular case feature (say genitive) will normally require this case feature to be overtly realized (by means of the preposition *de* or by means of the clitic *en*, which are alternative ways of realizing genitive case), but, if the realization of the oblique should cause a violation of \*P+CP, an alternative expression is chosen that does not cause this violation, even though it fails to realize the case requirement. This can be done in an OT framework, although other ways of obtaining preposition-less oblique clauses are possible.

In this way, eliminating COMP from the inventory of GFs not only results in a simplification of the framework (it is preferable to have fewer theoretical constructs), but also in a simplification of the analysis (verbs that alternate between taking a PP complement and a plain clausal complement, such as *convèncer*, have only one subcategorization frame, with an OBL, rather than two, one with an OBL and one with a COMP) and it reduces the redundancy in the theory (the c-structure realization of COMP is predictably clausal, i.e., CP or IP, but not NP or PP, whereas in the framework without COMP, both OBL and OBJ can map onto either a nominal or a clausal category) and makes it possible to state generalizations that are obscured in the framework with COMP (e.g., the fact that the clitic realization that corresponds to a clausal complement is the one that corresponds to the object or oblique complement of the verb).

### 3.1.3 Mixed subject-object properties

It is generally assumed that the single core argument of unaccusative verbs alternates between subject and object.<sup>25</sup> It can be shown that this argument sometimes has objecthood properties and sometimes has subjecthood properties. A paradox arises when we observe that this argument can have both types of properties in the same structure.

Evidence for assuming that the single core argument of unaccusative verbs can be expressed as an object is provided by the possibility of encoding this argument by means of the partitive clitic in those languages that have it, such as Catalan, French, and Italian. Since Perlmutter (1983), Rosen (1984), and Burzio (1986) for Italian, (also Alsina 1996b for Catalan), the claim is that this clitic must corre-

---

<sup>25</sup>The Unaccusative Hypothesis – the idea that intransitive verbs are classified into two classes depending on whether their core argument has some objecthood properties or not – was originally proposed by Perlmutter (1978) within the framework of Relational Grammar and subsequently adapted to other frameworks. See Section 2.3.2 for the different behavior of unaccusative and unergative verbs with respect to auxiliary selection in Italian.

spond to a direct object.<sup>26</sup> Example (39) shows that the unaccusative verb *sortir* ‘go out’ in Catalan allows its single core argument to be expressed by means of the partitive clitic, which, in this example, corresponds to the postverbal NP *un*. Given the claim just noted, this NP has to be a direct object.

(39) Catalan (Alsina & Yang 2018: 48)

Cada dia surten molts trens, però avui només n’ ha sortit un.  
 every day leave.3PL many trains but today only EN has left one  
 ‘Every day many trains leave, but today only one has left.’

Additional evidence supporting the claim that the argument partially encoded by the partitive clitic is an object comes from past participle agreement. In Catalan, the past participle optionally agrees in gender and number with a third person object clitic, when co-occurring with the perfective auxiliary *haver* ‘have’. The partitive clitic is one of the third person object clitics that can trigger past participle agreement, as in (40):

(40) Catalan (Fabra 1912: 160)

N’ han arribats molts.  
 EN have.PL arrive.PTCP.M.PL many.M.PL  
 ‘Many have arrived.’

In addition, the possibility of expressing the single direct argument of an intransitive verb as a bare indefinite NP provides further evidence for the objecthood of this argument, given the observation that this type of expression is excluded for the subject of transitive verbs.

Alongside the object encoding of the single direct argument in examples like (39)–(40), it is also possible for this argument to be expressed as the subject. The clearest evidence for this alternative encoding is the possibility of subject pro-drop. In a subject pro-drop language such as Catalan, a subject (and only a subject) can be null and be interpreted as having a definite referent, which indicates that, in (41), the missing argument, the logical subject of *sortir* ‘leave’, is its subject:

<sup>26</sup>The claim that, among intransitive verbs, only unaccusatives allow the partitive clitic, though commonly accepted, has been questioned by various scholars, who have pointed out that unergative verbs also allow the partitive clitic corresponding to their single core argument, at least under certain circumstances, such as Lonzi (1986) and Saccon (1995) for Italian, Cortés & Gavarró (1997) and Alsina & Yang (2018) for Catalan. Regardless of the correctness of this claim, the shared assumption is that the partitive clitic in these languages corresponds to a DO, which implies that the single core argument of an intransitive verb can be encoded as an object.

Alex Alsina

- (41) Catalan (based on Alsina & Yang 2018: 50)

Avui Ø surten tard.

today leave.3PL late

‘Today they are leaving late.’

If we should take verb agreement to be a subjecthood diagnostic in Catalan, we would have a problem in examples like (39)–(40). We find that the verb does not only agree with the subject, as is the case in (41), but also with the argument that is claimed to be an object. In (40), for example, the single core argument of *arribar* ‘arrive’ is expressed as the NP *molts* ‘many’, which has been argued to be an object, and yet this object agrees with the finite verb form *han*. But there is no need to assume that the agreement trigger is a subject. The verbal agreement facts of languages like Icelandic or Hindi indicate that the verb can agree with a grammatical function other than the subject, provided that it is in nominative case. And there is independent evidence that this is the case in Catalan as well. As shown in Alsina & Vigo (2014), in copular constructions with a predicative NP in Catalan, which are characterized by having two nominative phrases, the verb agrees with the nominative phrase that is higher in a person-number hierarchy where first and second person outrank third person and, among third persons, plural outranks singular (similar facts are found in Spanish and Italian). This indicates that what is necessary is for the agreement trigger to be a nominative expression.

Alsina & Yang (2018) propose an argument realization theory in which case is assigned to arguments independently of their GF and has the effect of constraining the GF assigned to an argument. According to their case assignment principles, nominative is assigned as a default to a core argument: A core argument that is not assigned dative or accusative case receives nominative. A constraint disallowing subjects with a case value other than nominative ensures that subjects in Catalan, and in the Romance languages in general, are nominative. Crucially, while all subjects are nominative, not all nominative arguments are subjects. The single core argument of an unaccusative verb is assigned nominative case and maps either onto the subject or the object.<sup>27</sup>

Thus, the paradox noted at the beginning of this subsection disappears. The single core argument of an unaccusative seems to have simultaneous subjecthood and objecthood properties: in examples like (39)–(40) it is encoded by the

<sup>27</sup> Alsina & Yang (2018) assume that this nominative argument maps onto the subject, when it is definite, and onto the object, when it is indefinite. This follows from treating the Subject Condition as a constraint in an OT setting and ranking it below an Indefinite Subject Ban, which penalizes an indefinite subject, in subject pro-drop languages like Catalan. So, the single core argument is a subject in an example like (41), but is an object in examples like (39)–(40).



partitive clitic and triggers past participle agreement, which are properties of objects, and it triggers finite verb agreement, which is usually assumed to be a property of subjects. However, once we observe that finite verb agreement is triggered by the nominative argument, all we need to assume is that the single core argument of a verb is always nominative and alternates between the subject and the object functions. As a nominative object, it has the standard objecthood properties, shared with accusative objects, and triggers finite verb agreement, a property of nominative arguments.

## 3.2 Passive and impersonal constructions

In this subsection we deal with passive and impersonal constructions. In Section 3.2.1, we compare the participial passive (or passive with auxiliary ESSE ‘be’) and the reflexive passive. And in Section 3.2.2, we review the evidence for considering the reflexive passive and the reflexive impersonal as the same or different constructions.

### 3.2.1 Two passive constructions

All Romance languages have two passive constructions, which we will call the participial passive and the reflexive passive. (The reflexive impersonal construction will be discussed in Section 3.2.2.) The participial passive is characterized by having the main predicate in the past participial form,<sup>28</sup> by the agreement in gender and number of this participle with its subject, by the fact that this subject has the same thematic role as the accusative object of the corresponding active form, and by the fact that the thematic role of the subject of the corresponding active form is either unexpressed or expressed by means of an oblique phrase (introduced by the preposition *da* in Italian, *par* in French, *por* in Spanish and Portuguese, *per* in Catalan, etc.). The passive participle can be used heading an adjunct clause, modifying either a clause or a noun, or as the main predicate of the clause along with a special auxiliary for passive clauses – the equivalent of *be* in the different languages (*ser*, *être*, *essere*, etc., although some languages have additional passive “auxiliaries,” such as *venire* or *andare* in Italian), as in the Catalan examples in (42):

<sup>28</sup>The assumption that past participles (of transitive verbs) can be passive and that it is the participial morphology that signals that the construction is passive is made in Bresnan (1982: 9–10) for English and in Loporcaro et al. (2004) for Romance, among others. The syntactic structure in which the participle is used (e.g., whether the auxiliary is ‘be’ or ‘have’) constrains the choice of the active or passive reading of the participle.

*Alex Alsina*

(42) Catalan

- a. Examinada la situació pels experts, la solució  
 examine.PTCP.F.SG the.F.SG situation by.the experts the solution  
 arribarà aviat.  
 arrive.FUT.3SG soon.  
 ‘Once the situation has been examined by the experts, the solution  
 will arrive soon.’
- b. La situació serà estudiada pels experts fins a  
 the.F.SG situation be.FUT.3SG study.PTCP.F.SG by.the experts until A  
 l’ últim detall.  
 the last detail  
 ‘The situation will be studied by the experts up to the last detail.’

Participial passives are also known as periphrastic passives, as they require an auxiliary in order to function as the main predicate of a clause other than an adjunct clause; however, since they can occur without an auxiliary in adjunct clauses such as in (42a), the term “participial passive” seems more appropriate.

The reflexive passive (or “Middle *se*” to use Grimshaw’s (1982) term) is characterized by the use of the reflexive clitic in the third person. The effects of this clitic on the mapping between arguments and GFs are very similar to those of the participial passive: the logical subject is suppressed, i.e., not expressed as a direct GF, and the direct object of the active form is the nominative GF, typically the subject. However, with the reflexive passive, the suppressed logical subject is generally not expressible as an oblique phrase. Morphologically, the reflexive passive is identical to the anaphoric and inherent uses of the reflexive reviewed in Section 2.2.3 and Section 2.3 and, potentially, gives rise to ambiguities with those uses of the reflexive. Two examples of reflexive passives in Catalan are given in (43), using verbs that, without the reflexive clitic, are transitive (i.e. take a direct, or accusative, object).

(43) Catalan

- a. Aquesta obra s’ estrenarà demà.  
 this play REFL premiere.FUT.3SG tomorrow  
 ‘This play will be premiered tomorrow.’
- b. Es preparen moltes pizzes en aquest local.  
 REFL prepare.3PL many pizzas in this establishment  
 ‘Many pizzas are prepared in this establishment.’

The direct object of the non-reflexive form corresponds to the nominative GF in the reflexive passive. As a nominative GF, it shows agreement with the verb: singular in (43a) vs. plural in (43b). It can be the subject, and often is (see Section 3.1.3): as such, it can appear in clause-initial position without an agreeing clitic on the verb, as in (43a), can be omitted with a definite interpretation, as in (44a), and cannot be expressed by means of a definite clitic, as in (44b):<sup>29</sup>

## (44) Catalan

- a. S' estrenarà demà.  
REFL premiere.FUT.3SG tomorrow  
'It will be premiered tomorrow.'
- b. \*Se les preparen en aquest local.  
REFL 3PL.F.ACC prepare.3PL in this establishment  
'They are prepared in this establishment.'

In subject pro-drop languages, like Catalan, subjects can be omitted with a definite interpretation, accounting for (44a). And definite object clitics such as *les* can only correspond to objects, which explains (44b).

Whereas the anaphoric and inherent uses of the reflexive clitic are compatible with all person features (first, second, and third), the reflexive passive can only occur with the third person clitic. It is not possible to have a reflexive passive with a first or second person subject, as that would require a first or second person reflexive clitic. Compare a well-formed participial passive with a first person subject, (45a), with the corresponding ill-formed reflexive passive, (45b).

## (45) Catalan

- a. He estat vist passant per la plaça.  
have.1SG been seen passing by the square  
'I have been seen walking across the square.'
- b. \*M' he vist passant per la plaça.  
me have.1SG seen passing by the square  
'I have been seen walking across the square.'

The two passive constructions are different morphologically, but share the definitional properties of a passive construction: the logical subject cannot be encoded as a direct GF and there is an internal argument encoded as a nominative GF, often the subject.

<sup>29</sup>(44b) is grammatical with an anaphoric interpretation, irrelevant here: 'They prepare them for themselves in this establishment.'

*Alex Alsina*

### 3.2.2 Reflexive passive and reflexive impersonal: one or two constructions?

The construction that we may call the impersonal reflexive, which is common at least in Spanish, Catalan, and Italian, like the reflexive passive also involves the reflexive clitic. It has a passive-like interpretation, as the argument that would be the subject without the reflexive clitic is unexpressed and interpreted as an arbitrary or unspecified human. It is found with intransitive predicates of both agentive and non-agentive types, as in (46). It also occurs with transitive verbs, in which case the internal argument should be analyzed as an accusative object because it does not agree with the verb and can be expressed by means of a definite object clitic, as in (47).

#### (46) Catalan

- a. Demà      no es      treballa.  
tomorrow not REFL work.3SG.  
'There is no work tomorrow.'
- b. No se      surt              fins que ho digui              jo.  
not REFL go.out.3SG until that HO say.SBJV.1SG I  
'No one goes out until I say so.'
- c. S'      ha              de ser      tossut      per fer      això.  
REFL have.3SG of be.INF stubborn to do.INF this  
'You've got to be stubborn to do this.'

#### (47) Catalan

- a. S'      ha              seguit      els sospitosos fins al      seu pis.  
REFL have.3SG followed the suspects until A.the their flat  
'The suspects have been followed up to their flat.'
- b. Se' ls              ha              seguit      fins al      seu pis.  
REFL 3PL.M.ACC have.3SG followed until A.the their flat  
'They have been followed up to their flat.'

There are clear similarities between the reflexive passive and the impersonal reflexive constructions that make it desirable to assume that the reflexive clitic performs the same function in both cases. The two constructions share the fact that the logical subject is not expressed and is interpreted as an arbitrary or unspecified human and that they can only be used with the third person form of the reflexive clitic. For this reason it is not possible to distinguish them semanti-

cally. This has led some researchers, such as Cardona (2015), to claim that both constructions should be treated as a passive construction.<sup>30</sup>

However, no attempt to derive the two constructions from a single operation performed by the reflexive clitic has successfully explained all the facts of both constructions. The main objections to such a reductionist approach, which would assume that the reflexive clitic is the morphological exponent of a passive operation in both constructions, have been pointed out in Yang (2019). The first objection concerns the conditions on accusative case assignment. Accusative case can only be assigned in an argument structure that contains an external argument expressed as a direct function. This explains the observation that passive sentences in Romance, including reflexive passive sentences, do not have accusative objects: for this reason the reflexive passive (44b) is ungrammatical, as it has an object clitic that corresponds to an accusative object. But if the impersonal reflexive were also a passive form, we would not be able to explain the grammaticality of (47b), which does contain a clitic corresponding to an accusative object. As a passive form, it would not have a direct function mapped onto the external argument and accusative case should not be assigned.

The second objection has to do with the observation that the impersonal *se* can occur in constructions in which one cannot argue that a logical subject is being suppressed, either because the argument that is interpreted as a generic or arbitrary human is not a thematic argument of the predicate or because it is not the logical subject. This is arguably the situation with copular sentences, such as (46c), on the assumption that the subject of the copula is not an argument of the copula, but of its predicative complement. And it is definitely the case when impersonal *se* is attached to a participial passive sentence, as in (48), from Yang (2019). Although such examples are rare and hard to contextualize, they are not ungrammatical.

- (48) Catalan (Institut d'Estudis Catalans 2016: 895)  
 Passava això quan s' era expulsat del partit.  
 happened this when REFL was expelled from.the party  
 'This is what happened when one was expelled from the party.'

The reflexive clitic cannot be the exponent of the suppression of the logical subject of the verb in participial form, because this argument is already suppressed

---

<sup>30</sup>See Bentley (2006) for an attempt to capture both the differences and the commonalities between the anaphoric, passive and impersonal uses of the reflexive clitic in Italian, within the framework of Role and Reference Grammar.

*Alex Alsina*

by the participial morphology. If anything is suppressed by the reflexive morphology, it is the subject of the copula, a non-thematic GF of this verb that controls the subject of the participial verb, which is not its logical subject.

Given these two objections to the unified analysis of the reflexive passive and the reflexive impersonal, it seems necessary to assume that they are two different constructions, as concluded in Yang (2019): The reflexive passive is a passive construction, in which the logical subject is suppressed, whereas the reflexive impersonal licenses a null, 3rd person singular subject, with an arbitrary human interpretation. This is also the proposal in Kelling (2006).

The reflexive passive and the reflexive impersonal, although different constructions, are in competition. According to Aranovich (2009), with dyadic predicates, in Spanish, the choice between the two constructions is determined by the animacy features of the internal argument. If this argument is animate, the reflexive impersonal construction is employed, but if it is inanimate the reflexive passive is preferred:<sup>31</sup>

(49) Spanish (Aranovich 2009: 623–624)

- a. Ayer se atrapó a los ladrones.  
yesterday REFL caught.3SG A the thieves  
'The thieves were caught yesterday.'
- b. Ayer se atraparon las pelotas  
yesterday REFL caught.3PL the balls  
'Yesterday, the balls were caught.'

(50) Spanish (Aranovich 2009: 623–624)

- a. \*Ayer se atraparon los ladrones.  
yesterday REFL caught.3PL the thieves  
'The thieves were caught yesterday.'
- b. \*Ayer se atrapó las pelotas  
yesterday REFL caught.3SG the balls  
'Yesterday, the balls were caught.'

Aranovich (2009) develops an analysis using Optimality Theory (OT) and Lexical Mapping Theory (LMT). In this analysis, the alternation between the reflexive impersonal and the reflexive passive is the result of a conflict between two

<sup>31</sup>While there might be a strong preference for the choice between the two constructions to depend on the animacy of the internal argument, sentences such as (50) are generally not considered to be ungrammatical.

constraints, one favoring the assignment of the subject function to the reflexive clitic and another one penalizing inanimate objects. The difference between the two constructions is reflected in the GF assigned to the reflexive clitic, which is a subject in the reflexive impersonal and an oblique in the reflexive passive. The reflexive passive avoids the marked configuration of an inanimate object by allowing the inanimate internal argument to be realized as the subject. See Aranovich (2009) for the details of the analysis.

## 4 Complex predicates

Complex predicates have been the object of investigation within LFG in a variety of languages since work such as Mohanan (1990, 1994), Matsumoto (1992), Alsina (1993, 1996b), and Butt (1993, 1995). For present purposes we can follow Butt's (1995: 2) definition and take a complex predicate to be a construction whose argument structure is complex, in the sense that two or more semantic heads contribute to it, and whose GF structure is that of a simple predicate. The Romance languages have made a significant contribution to this investigation, as they have several constructions that are analyzed as complex predicates, particularly, the causative construction and restructuring constructions. In Section 4.1 we examine the facts of these constructions and, in Section 4.2, we review some of the analyses that have been proposed for them.

### 4.1 The causative and restructuring constructions

#### 4.1.1 The causative construction

In contrast with languages where causative verb forms are a single word consisting of a stem and a causative affix (as in Chicheŵa and many Bantu languages), causative constructions in the Romance languages comprise two verb forms (leaving aside the fact that they can also be accompanied by auxiliaries): the causative verb and an infinitive complement.<sup>32</sup> There are two causative verbs,

<sup>32</sup>The Romance languages also include many verbs that are causative in meaning but cannot be considered to be complex predicates in the sense intended here as they are not decomposable into a base predicate and a causative predicate (whether bound morpheme or independent word). This is the case of *romper* 'break' or *abrir* 'open' in Spanish, or *chiudere* 'close' or *raffreddare* 'cool' in Italian. Some of these verbs, including the examples given, undergo the causative-anticausative alternation, which is signaled morphologically by means of the reflexive clitic on the anticausative member of the alternation (e.g. *romperse* or *abrirse* in Spanish and *chiudersi* or *raffreddarsi* in Italian). It is, therefore, an *anticausative* alternation, in Haspelmath's (1993) terms (see also Cennamo 2016: 971), in contrast with the Bantu pattern, where the alternation is causative.

Alex Alsina

which behave alike in most respects syntactically: *fare* ‘make’ and *lasciare* ‘let’ in Italian, and the corresponding pairs in French (*faire* and *laisser*), Spanish (*hacer* and *dejar*) or Catalan (*fer* and *deixar*),<sup>33</sup> see exx. (51)–(54).

What distinguishes the causative construction in Romance from other constructions in which a verb takes an infinitival complement is what we might call the monoclausality of the causative construction, that is, the fact that the causative verb and the infinitive behave as if they were part of one and the same clause from the point of view of the f-structure. As shown in Alsina (1997), the causative verb and the infinitive are a unit at the level of f-structure, very much like causative verbs in Chicheŵa, but are clearly two different units (i.e., two separate verbs) at the level of c-structure, unlike causative verb forms in Chicheŵa, which are a unit at both levels.

Following is some of the evidence in favor of the monoclausality of the causative construction:

#### 4.1.1.1 The case alternation on the causee

(I use the term causee here to refer to the logical subject of the infinitive, or embedded predicate, in the causative construction.). As shown in Section 2.3.1, example (22), repeated here as (51), the case of the causee depends on the transitivity of the embedded predicate: it is dative if the embedded predicate has an accusative object, and it is accusative otherwise.

(51) French (Grimshaw 1990: 153)

- a. Il fera                    boire      un peu de vin    \*(à) son enfant.  
    he make.FUT.3SG drink.INF a    bit    of wine    A his child  
    ‘He will make his child drink a little wine.’
- b. Il fera                    partir      {les/\*aux} enfants.  
    he make.FUT.3SG leave.INF the/\*A.the children  
    ‘He will make the children leave.’

This case alternation would be unexpected if the infinitive were the f-structure head of an embedded clause. By viewing the two verbs in the construction as forming a unit, a PRED, at f-structure, this case alternation can be made to follow

<sup>33</sup>Some of these verbs also admit a biclausal raising-to-object construction, in which both the causative verb and the dependent infinitive head their own clause and the object of the causative verb functionally controls the subject of the infinitival clause. This is the case of a French example such as *Elle a laissé Jean laver la voiture* ‘She let John wash the car.’ Since these constructions are not complex predicates, they will not be discussed here.



from a theory of argument realization in which dative case is assigned only as a marked option, that is, to the more prominent of two internal arguments (as proposed in Alsina 1996b and Alsina & Yang 2018).

#### 4.1.1.2 Clitic climbing

Clitics that correspond to argument roles of the embedded predicate usually appear attached to the causative verb (or to a higher auxiliary or restructuring verb), as in (52):

(52) Catalan

- a. Això m' hi ha fet pensar.  
that me HI has made think.INF  
'That made me think about it.'
- b. Aquests documents, els faré enquadernar.  
these documents 3PL.M.ACC I.will.make bind.INF  
'These documents, I will have them bound.'

The clitic *hi* in (52a) corresponds to the oblique complement of *pensar* 'think' and yet appears attached to the auxiliary of the causative verb; likewise in (52b), where the clitic *els* corresponds to the accusative object of *enquadernar* 'bind'. This property is not found with verbs that take an infinitival clausal complement, such as *semblar* 'seem', *caldre* 'be necessary', *convenir* 'be convenient', *insistir* 'insist, etc., in which case the clitics dependent on the infinitive appear attached to the infinitive.

#### 4.1.1.3 Reflexivization

The reflexive clitic can encode the binding of the logical subject of the causative predicate and an argument of the embedded predicate, as in (53a).

#### 4.1.1.4 Reflexive passive

A reflexive passive of the causative predicate, encoded by the reflexive clitic, can have an argument of the embedded predicate as its nominative argument, agreeing with the causative verb (or a higher auxiliary or restructuring verb), as in (53b).

(53) Catalan

*Alex Alsina*

- a. S' ha fet criticar durament.  
REFL has made criticize.INF hard  
'She has got herself criticized severely.'
- b. S' han fet arreglar les façanes del carrer principal.  
REFL have.3PL made fix.INF the façades of.the street main  
'The façades of the main street have been made to be repaired.'

#### 4.1.1.5 Passivization

Some Romance languages allow participial passivization of the causative construction, in which an argument of the embedded predicate is the subject of the passivized causative structure. This possibility is illustrated for Italian in (54a), from Frank (1996), whereas French is a language that does not allow it.

#### 4.1.1.6 Past participle agreement

Among those Romance languages in which the past participle of compound tenses agrees with the accusative object expressed as a clitic (or, depending on the language, in other cases as well), Italian has this phenomenon in causative constructions, as in (54b), although French does not.

#### (54) Italian

- a. Questo libro è stato fatto leggere a Mario da Giovanni.  
this book is been made read.INF A Mario by Giovanni  
'This book has been made to be read by Mario by Giovanni.'
- b. Le tavole, le ho fatte riparare a  
the.F.PL table.F.PL 3PL.F.ACC have.1SG make.PTCP.F.PL repair.INF A  
Gianni.  
Gianni  
'The tables, I have made Gianni repair them.'

Other phenomena that support the monoclausal treatment of the causative construction include *tough* movement, which in Romance is a clause-bound phenomenon: as it can affect the object of the embedded predicate in a causative construction, it shows that the causative predicate and the embedded predicate constitute a single complex predicate. Although the facts are quite compelling in this respect, there are some attempts to explain them adopting a biclausal approach, as in Yates (2002).

### 4.1.2 The restructuring construction

The restructuring construction, present in many of the Romance languages, but absent in modern French, is similar to the causative construction in that it also involves two verbs (not counting auxiliaries) that form a complex predicate and behave as if they belonged to the same clause, but differs from it in not increasing the valence of the embedded predicate. The list of restructuring verbs varies somewhat from language to language, and even from one speaker to another, but it typically includes verbs such as (using Catalan for the examples) *voler* ‘want’, *poder* ‘can, be able’, *saber* ‘know’, *venir a* ‘come to’, *anar a* ‘go to’, *tornar* ‘do again’, *començar a* ‘begin’, *acabar de* ‘finish’, etc. The construction was first described by Aissen & Perlmutter (1976) and Rizzi (1976),<sup>34</sup> who proposed an optional process of clause union or restructuring, respectively, in order to explain that a restructuring verb, such as those just mentioned, and a dependent verb can behave as if they were a single verb from the point of view of their GFs.

As with the causative construction, one of its salient features is the possibility of clitic climbing. Reflexivization and the reflexive passive are also possible with the restructuring construction. Some verbs allow participial passive and languages that have past participle agreement with the object in compound tenses also exhibit this phenomenon in the restructuring construction. In languages that have auxiliary selection, like Italian, the choice of auxiliary is determined by the embedded verb. To illustrate just some of these phenomena in Italian, *dovere* ‘have to’, as a verb taking an infinitival phrase, allows clitic climbing, as the position of the clitic *gli* illustrates in (55), and also allows, but does not require, the choice of auxiliary to be determined by the infinitive, as shown in (56):

(55) Italian (Rizzi 1982: 4).

- a. Gianni ha dovuto parlargli personalmente.  
Gianni has had.to speak.3SG.M.DAT personally
- b. Gianni gli ha dovuto parlare personalmente.  
Gianni 3SG.M.DAT has had.to speak personally  
‘Gianni has had to speak with him personally.’

(56) Italian (Rizzi 1982: 19).

- Piero ha / è dovuto venire con noi.  
Piero has / is had.to come with us  
‘Piero has had to come with us.’

<sup>34</sup> Although these works are better known through later publications, specifically Aissen & Perlmutter (1983) and Rizzi (1982), the fact that the first version of these works has the same date of publication suggests that they were developed independently of each other.

*Alex Alsina*

Interestingly, when clitic climbing takes place from an infinitive such as *venire* ‘come’, which selects *essere*, the option of using the *avere* auxiliary disappears, as shown in (57):

- (57) Italian (Rizzi 1982: 21)
- a. Maria c’ è dovuta venire molte volte.  
     Maria CL is had.to.F.SG come many times
  - b. \*?Maria ci ha dovuto venire molte volte.  
     Maria CL has had.to come many times  
     ‘Maria has had to come there many times.’

Restructuring is optional, accounting for the options in (55)–(56). When restructuring occurs, clitic climbing is required and auxiliary choice is determined by the dependent infinitive, which accounts for the contrast in (57).

#### 4.2 Analyses of the Romance complex predicates

Alsina (1996b), adapting Alsina’s (1992) proposal for the causative predicate in Chicheŵa, assumes that the causative predicate in Romance has a three-place argument structure, in which there is a causer, an affected (or acted-upon) argument, and a caused event. In addition, the affected argument is fused with an argument of the caused event, so that there is a GF that corresponds to two argument roles: the affected argument of the causative predicate and another role of the caused event. The caused event position in the causative argument structure is filled by the predicate of the infinitive in the causative construction.

In this way, the causative complex predicate is formed in the syntax in Romance, whereas it is formed in the lexicon in Chicheŵa. As argued in Alsina (1997), the causative complex predicate is the same in the two languages as far as the argument structure is concerned, but they differ in that it corresponds to a single word in Chicheŵa (containing a verb stem and a causative suffix), but it corresponds to two words in Romance (the causative verb and the infinitive). If the lexicon is the linguistic component in which words are formed, as well as stored, and the syntax operates with fully formed words, the difference between the two languages concerning causative predicates resides in the component in which this complex predicate is formed: the lexicon in Chicheŵa, the syntax in Romance. Given that this proposal implies some departure from classical LFG assumptions (such as the idea that the list of GFs that a predicate requires is fixed in the lexicon and cannot be altered in the syntax), there are alternative proposals that assume that the causative complex predicate is formed in the lexicon, as in

Frank (1996), in spite of the fact that it corresponds to two distinct words in the syntax.

The treatment of causatives in Alsina (1996b) can be adapted to handle restructuring constructions. The only difference is that a restructuring verb either takes an event argument as its sole argument, as would be the case of *dovere*, or takes an additional argument role that is fused with the logical subject of the event argument, as would be the case of *volere* ‘want’ or *venire* ‘come’. In either case, the resulting restructuring construction has no more expressed arguments than the base predicate, the infinitive. When restructuring takes place, the auxiliary selection properties of the construction are determined by the base predicate and the highest verb in the sequence of restructured verbs, including auxiliaries, is the one to which clitics are attached.

The idea that predicate formation may take place in the syntax, as opposed to the lexicon, has been met with some resistance by some LFG practitioners. Yet, the alternative, namely, that complex predicate formation with restructuring and causative verbs takes place in the lexicon, is hard to maintain given that the sequence of such verbs is potentially unlimited. Following are two examples with a long sequence of restructuring and causative verbs in Catalan:

(58) Catalan

- a. La            va            haver    de tornar    a començar a escriure.  
       3SG.F.ACC PAST.3SG have.INF to repeat.INF to begin.INF to write.INF  
       ‘She had to start writing it again.’
- b. L’            hi        he            volgut    fer            acabar    de recitar.  
       3SG.M.ACC 3.DAT have.1SG want.PTCP make.INF finish.INF of recite.INF  
       ‘I wanted to make him finish reciting it.’

In both examples the clitics (*la* in (58a) and *l’hi* in (58b)) are thematically related to the base predicate, but appear attached to the matrix verb (the past tense auxiliary *va* in (58a) and the perfective auxiliary *he* in (58b)), indicating that there is complex predicate formation involving all the verbs from the auxiliary to the base predicate.

An issue that Alsina (1996b, 1997) does not address is how the light verb (the causative, restructuring, or auxiliary verb) in a complex predicate imposes form requirements on the dependent verb. Some verbs, such as the causative verbs and restructuring verbs like *poder* ‘can’ and *voler* ‘want’, require a prepositionless infinitive, as seen in (58b) and preceding examples. Other verbs require a specific preposition before the infinitive: *haver* in (58a) and *acabar* in (58b) require

Alex Alsina

the preposition *de* before the infinitive; *tornar* and *començar* in (58a) require the preposition *a* before the infinitive.

The traditional LFG way to capture these dependencies is through the f-structure. However, if the f-structure is “flat” so that there is no feature structure corresponding to the dependent verb that is distinct from that of the embedding verb, this mechanism is no longer available. Andrews & Manning (1999) notice this problem and propose a way to capture the monoclausality of complex predicates, while retaining an embedding relation between the light verb and its dependent verb. The leading idea in Andrews & Manning (1999) is that the features traditionally assumed to be part of f-structure are grouped into three classes:  $\rho$ : grammatical relations (SUBJ, OBJ, ...);  $\alpha$ : argument structure features such as PRED and others; and  $\mu$ : morphosyntactic features (GEND, NUM, TENSE, etc.). In addition, every node in the c-structure specifies which of these feature classes is shared with its mother node. In this way, it is possible to achieve a flat f-structure as far as GFs are concerned by having the two verbs in the complex predicate share the  $\rho$  class with the mother, but having only the light verb share its  $\alpha$  and  $\mu$  features with the mother, whereas the dependent verb would contribute its  $\alpha$  and  $\mu$  features to an ARG attribute. ARG is not a grammatical relation, but one of the features on the  $\alpha$ -projection. Having this ARG feature allows the light verb to specify form features on its dependent verb (whether it is an infinitive or a gerund, what preposition it requires, if any, etc.). The embedding at the  $\alpha$ -projection allows Andrews & Manning (1999) to capture the fact that the order of the light verbs is reflected in the meaning of the complex predicate, as in the following Catalan examples:

(59) Catalan (Alsina 1997)

- a. Li            acabo de fer            llegir    la carta.  
               3SG.DAT I.finish of make.INF read.INF the letter  
               ‘I finish making him read the letter.’ or ‘I just made him read the letter.’
- b. Li            faig acabar    de llegir    la carta.  
               3SG.DAT I.make finish.INF of read.INF the letter  
               ‘I make him finish reading the letter.’

This proposal is not very different from the proposal in Butt et al. (1996), which is designed to account for structures with auxiliaries, but can easily be applied to the analysis of complex predicates. Butt et al. (1996) propose to split the traditional f-structure into two structures, or projections: the grammatical features of verb forms (having to do with whether the form is an infinitive, a gerund, etc.) are removed from the f-structure and placed in the m-structure, which allows the f-structure of an auxiliated structure, and of complex predicates, to be “flat”,

i.e., not containing an embedding relation between the auxiliary or restructuring verb and its dependent verb. The dependent verbs in auxiliated structures, and by extension in complex predicates, provide their form features to a DEP attribute. In this way, the auxiliary, or the light, verb can impose form requirements on their DEP (the dependent verb) achieving a similar result to that achieved by Andrews & Manning (1999). More recent LFG developments in the analysis of complex predicates include Andrews (2007), Homola & Coler (2013), and Lowe (2016), which shift the burden of explanation onto the semantics.

## Acknowledgments

I am very grateful to Nigel Vincent and two anonymous reviewers for their comments, which have helped improve this chapter in many ways.

## References

- Aissen, Judith L. & David M. Perlmutter. 1976. Clause reduction in Spanish. In *Proceedings of the 2nd annual meeting of the Berkeley Linguistics Society*, 1–30. Berkeley: Berkeley Linguistics Society. DOI: 10.3765/bls.v2i0.2283.
- Aissen, Judith L. & David M. Perlmutter. 1983. Clause reduction in Spanish. In David M. Perlmutter (ed.), *Studies in Relational Grammar 1*, 360–403. Chicago: University of Chicago Press. Earlier version published as Aissen & Perlmutter (1976).
- Alencar, Leonel F. de & Carmen Kelling. 2005. Are reflexive constructions transitive or intransitive? Evidence from German and Romance. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '05 conference*, 1–20. Stanford, CA: CSLI Publications.
- Alsina, Alex. 1992. On the argument structure of causatives. *Linguistic Inquiry* 23. 517–555.
- Alsina, Alex. 1993. *Predicate composition: A theory of syntactic function alternations*. Stanford, CA: Stanford University. (Doctoral dissertation).
- Alsina, Alex. 1996a. Passive types and the theory of object asymmetries. *Natural Language & Linguistic Theory* 14. 673–723. DOI: 10.1007/bf00133361.
- Alsina, Alex. 1996b. *The role of argument structure in grammar: Evidence from Romance* (CSLI Lecture Notes). Stanford, CA: CSLI Publications.
- Alsina, Alex. 1997. Causatives in Bantu and Romance. In Alex Alsina, Joan Bresnan & Peter Sells (eds.), *Complex predicates*, 203–246. Stanford, CA: CSLI Publications.

Alex Alsina

- Alsina, Alex. 2001. On the nonsemantic nature of argument structure. *Language Sciences* 23(4-5). 355–389. DOI: 10.1016/s0388-0001(00)00030-9.
- Alsina, Alex. 2010. The Catalan definite article as lexical sharing. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 5–25. Stanford, CA: CSLI Publications.
- Alsina, Alex. 2011. Tres classes de determinants en català. In Maria-Rosa Lloret & Clàudia Pons (eds.), *Noves aproximacions a la fonologia i la morfologia del català*, 11–36. Alacant: Institut Interuniversitari de Filologia Valenciana.
- Alsina, Alex. 2020. Obligatory clitic expression, clitic omission, and the morphology-syntax interface. In Miriam Butt & Ida Toivonen (eds.), *Proceedings of the LFG '20 conference*, 5–25. Stanford, CA: CSLI Publications.
- Alsina, Alex & Sam A. Mchombo. 1990. The syntax of applicatives in Chichewa: Problems for a theta theoretic asymmetry. *Natural Language & Linguistic Theory* 8(4). 493–506. DOI: 10.1007/bf00133691.
- Alsina, Alex, K. P. Mohanan & Tara Mohanan. 2005. How to get rid of the COMP. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '05 conference*, 21–41. Stanford, CA: CSLI Publications.
- Alsina, Alex & Eugenio M. Vigo. 2014. Copular inversion and non-subject agreement. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 5–25. Stanford, CA: CSLI Publications.
- Alsina, Alex & Fengrong Yang. 2018. Catalan intransitive verbs and argument realization. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '18 conference*, 46–66. Stanford, CA: CSLI Publications.
- Andrews, Avery D. 1990. Unification and morphological blocking. *Natural Language & Linguistic Theory* 8(4). 507–557. DOI: 10.1007/bf00133692.
- Andrews, Avery D. 2007. Projections and glue for clause-union complex predicates. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '07 conference*, 44–65. Stanford, CA: CSLI Publications.
- Andrews, Avery D. & Christopher D. Manning. 1999. *Complex predicates and information spreading in LFG*. Stanford, CA: CSLI Publications.
- Aranovich, Raúl. 2009. Feature-based argument mapping and animacy optimization in impersonal passives. *Linguistics* 47(3). 619–652. DOI: 10.1515/ling.2009.021.
- Aranovich, Raúl. 2012. A Lexical-Functional account of Spanish dative usage. In Monique Lamers & Peter de Swart (eds.), *Case, word order and prominence*, 17–41. Dordrecht: Springer.
- Baker, Mark C. 1988a. *Incorporation: A theory of grammatical function changing*. Chicago: University of Chicago Press.



- Baker, Mark C. 1988b. Theta theory and the syntax of applicatives in Chicheŵa. *Natural Language & Linguistic Theory* 6. 353–389. DOI: [10.1007/bf00133903](https://doi.org/10.1007/bf00133903).
- Barbu, Roxana-Maria & Ida Toivonen. 2018. Romanian object clitics: Grammaticalization, agreement and lexical splits. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '18 conference*, 67–87. Stanford, CA: CSLI Publications.
- Barron, Julia. 2000. The morphosyntactic correlates of finiteness. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '00 conference*, 25–43. Stanford, CA: CSLI Publications.
- Belyaev, Oleg, Anastasia Kozhemyakina & Natalia Serdobolskaya. 2017. In defense of COMP: Complementation in Moksha Mordvin. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '17 conference*, 83–103. Stanford, CA: CSLI Publications.
- Bentley, Delia. 2006. *Split intransitivity in Italian*. Berlin/New York: De Gruyter Mouton. DOI: [10.1515/9783110896053](https://doi.org/10.1515/9783110896053).
- Bonet, Eulàlia. 1991. *Morphology after syntax: Pronominal clitics in Romance*. Distributed by MIT Working Papers in Linguistics. Cambridge, MA: Massachusetts Institute of Technology. (Doctoral dissertation).
- Bonet, Eulàlia. 1995. Feature structure of Romance clitics. *Natural Language & Linguistic Theory* 13(4). 607–647. DOI: [10.1007/bf00992853](https://doi.org/10.1007/bf00992853).
- Bresnan, Joan. 1982. The passive in lexical theory. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 3–86. Cambridge, MA: The MIT Press.
- Bresnan, Joan & Sam A. Mchombo. 1987. Topic, pronoun, and agreement in Chicheŵa. *Language* 63. 741–782. DOI: [10.2307/415717](https://doi.org/10.2307/415717).
- Bresnan, Joan & Lioba Moshi. 1990. Object asymmetries in comparative Bantu syntax. *Linguistic Inquiry* 21. 147–186.
- Burzio, Luigi. 1986. *Italian syntax: A Government-Binding approach*. Dordrecht: Reidel.
- Butt, Miriam. 1993. *The structure of complex predicates in Urdu*. Stanford, CA: Stanford University. (Doctoral dissertation). <https://ojs.ub.uni-konstanz.de/jsal/dissertations/diss-butt.pdf>.
- Butt, Miriam. 1995. *The structure of complex predicates in Urdu* (Dissertations in Linguistics). Stanford, CA: CSLI Publications.
- Butt, Miriam, María-Eugenia Niño & Frederique Segond. 1996. Multilingual processing of auxiliaries in LFG. In D. Gibbon (ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS conference*, 111–122. Berlin: Mouton de Gruyter.

Alex Alsina

- Cardinaletti, Anna. 2004. Toward a cartography of subject positions. In Luigi Rizzi (ed.), *The structure of CP and IP: The cartography of syntactic structures*, vol. 2, 115–165. Oxford: Oxford University Press.
- Cardinaletti, Anna & Lori Repetti. 2010. Proclitic vs enclitic pronouns in northern Italian dialects and the null-subject parameter. In Roberta D’Alessandro, Adam Ledgeway & Ian Roberts (eds.), *Syntactic variation: The dialects of Italy*, 119–134. Cambridge, UK: Cambridge University Press.
- Cardona, Margrete. 2015. La enseñanza de las construcciones pasivas e impersonales con *se* en *E/LE*. ¿Cuántas distinciones son necesarias? *Didáctica. Lengua y Literatura* 27. 73–96.
- Carretero García, Paloma. 2017. Agreement in Asturian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’17 conference*, 188–208. Stanford, CA: CSLI Publications.
- Carretero García, Paloma. 2018. Dative arguments in psychological predicates in Spanish. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’18 conference*, 150–170. Stanford, CA: CSLI Publications.
- Cennamo, Michela. 2016. Voice. In Adam Ledgeway & Martin Maiden (eds.), *The Oxford guide to the Romance languages*, 967–980. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199677108.003.0060.
- Colomina i Castanyer, Jordi. 2002. Paradigmes flectius de les altres classes nominals. In Joan Solà, Maria Rosa Lloret, Joan Mascaró & Manuel Pérez Saldanya (eds.), *Gramàtica del català contemporani*, vol. 1, 535–582. Barcelona: Editorial Empúries.
- Cortés, Corinne & Anna Gavarró. 1997. Subject-object asymmetries and the clitic *en*. In James R. Black & Virginia Motapanyane (eds.), *Clitics, pronouns and movement*, 39–62. Amsterdam: John Benjamins.
- Crysmann, Berthold. 1997. Cliticization in European Portuguese using parallel morpho-syntactic constraints. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’97 conference*, 1–14. Stanford, CA: CSLI Publications.
- Dalrymple, Mary & Helge Lødrup. 2000. The grammatical functions of complement clauses. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’00 conference*, 104–121. Stanford, CA: CSLI Publications.
- Dragomirescu, Adina & Alexandru Nicolae. 2016. Case. In Adam Ledgeway & Martin Maiden (eds.), *The Oxford guide to the Romance languages*, 911–923. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199677108.003.0056.
- Estigarribia, Bruno. 2005. Direct object clitic doubling in OT-LFG: A new look at Rioplatense Spanish. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’05 conference*, 116–135. Stanford, CA: CSLI Publications.

- Estigarribia, Bruno. 2013. Rioplatense Spanish clitic doubling and “tripling” in Lexical-Functional Grammar. In Chad Howe, Sarah E. Blackwell & Margaret Lubbers Quesada (eds.), *Selected proceedings of the 15th Hispanic Linguistics Symposium*, 297–309. Somerville, MA: Cascadilla Proceedings Project.
- Fabra, Pompeu. 1912. *Gramàtica de la lengua catalana*. Barcelona: L’Avenç.
- Falk, Yehuda N. 2001. *Lexical-Functional Grammar: An introduction to parallel constraint-based syntax*. Stanford, CA: CSLI Publications.
- Fassi Fehri, Abdelkader. 1984. Agreement in Arabic, binding and coherence. Presented at the Conference on Agreement in Natural Language, Stanford University.
- Fassi Fehri, Abdelkader. 1988. Agreement in Arabic, binding and coherence. In Michael Barlow & Charles A. Ferguson (eds.), *Agreement in natural language*, 107–158. Stanford, CA: CSLI Publications.
- Fernández-Soriano, Olga. 1999. Two types of impersonal sentences in Spanish: Locative and dative subjects. *Syntax* 2(2). 101–140. DOI: 10.1111/1467-9612.00017.
- Fischer, Susann. 2002. *The Catalan clitic system: A diachronic perspective on its syntax and phonology*. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110892505.
- Fontana, Josep M. 1993. *Phrase structure and the syntax of clitics in the history of Spanish*. Philadelphia: University of Pennsylvania. (Doctoral dissertation).
- Fontana, Josep M. 1996. Phonology and syntax in the interpretation of the Tobler-Mussafia law. In Aaron L. Halpern & Arnold M. Zwicky (eds.), *Approaching second: Second position clitics and related phenomena*, 41–82. Stanford, CA: CSLI Publications.
- Forst, Martin. 2006. comp in (parallel) grammar writing. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’06 conference*, 222–239. Stanford, CA: CSLI Publications.
- Frank, Anette. 1996. A note on complex predicate formation: Evidence from auxiliary selection, reflexivization, passivization and past participle agreement in French and Italian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’96 conference*. Stanford, CA: CSLI Publications.
- Gazdik, Anna. 2008. French interrogatives in an OT-LFG analysis. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’08 conference*, 272–290. Stanford, CA: CSLI Publications.
- Gazdik, Anna. 2010. Multiple questions in French and in Hungarian: An LFG account. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’10 conference*, 249–269. Stanford, CA: CSLI Publications.

Alex Alsina

- Grimshaw, Jane. 1982. On the lexical representation of Romance reflexive clitics. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 87–148. Cambridge, MA: The MIT Press.
- Grimshaw, Jane. 1990. *Argument structure*. Cambridge, MA: The MIT Press.
- Grimshaw, Jane. 1997. The best clitic: Constraint conflict in morphosyntax. In Liliane Haegeman (ed.), *Elements of grammar: Handbook in generative syntax*, 169–196. Dordrecht: Kluwer Academic Publishers. DOI: 10.1007/978-94-011-5420-8\_4.
- Halle, Morris & Alec Marantz. 1993. Distributed morphology and the pieces of inflection. In Kenneth Hale & Samuel Jay Keyser (eds.), *The view from Building 20: Essays in linguistics in honor of Sylvain Bromberger*, 111–176. Cambridge, MA: The MIT Press.
- Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In Bernard Comrie & Maria Polinsky (eds.), *Causatives and transitivity*, 87–120. Amsterdam: John Benjamins. DOI: 10.1075/slcs.23.05has.
- Homola, Petr & Matt Coler. 2013. Causatives as complex predicates without the restriction operator. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '13 conference*, 316–334. Stanford, CA: CSLI Publications.
- Institut d'Estudis Catalans. 2016. *Gramàtica de la llengua catalana*. Barcelona: Institut d'Estudis Catalans.
- Katerinov, Katerin. 1975. *La lingua italiana per stranieri: Corso medio*. Perugia: Edizioni Guerra.
- Kelling, Carmen. 2006. Spanish *se*-constructions: The passive and the impersonal construction. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '06 conference*, 275–288. Stanford, CA: CSLI Publications.
- Ledgeway, Adam & Martin Maiden (eds.). 2016. *The Oxford guide to the Romance languages*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199677108.001.0001.
- Lødrup, Helge. 2004. Clausal complementation in Norwegian. *Nordic Journal of Linguistics* 27(1). 61–95. DOI: 10.1017/s0332586504001155.
- Lødrup, Helge. 2012. In search of a nominal COMP. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 383–404. Stanford, CA: CSLI Publications.
- Lonzi, Lidia. 1986. Pertinenza della struttura tema-rema per l'analisi sintattica. In Harro Stammerjohann (ed.), *Tema-rema in italiano. Theme-rheme in Italian. Thema-Rhema im Italienischen*, 99–120. Tübingen: Narr.
- Loporcaro, Michele. 2016. Auxiliary selection and participial agreement. In Adam Ledgeway & Martin Maiden (eds.), *The Oxford guide to the Romance lan-*

- guages, 802–818. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199677108.003.0049.
- Loporcaro, Michele, Lorenza Pescia & Maria Ana Ramos. 2004. Costrutti dipendenti participiali e participi doppi in portoghese. *Revue de linguistique romane* 68. 15–46.
- Lowe, John J. 2016. Complex predicates: An LFG+glue analysis. *Journal of Language Modelling* 3. 413–462. DOI: 10.15398/jlm.v3i2.125.
- Luís, Ana & Ryo Otaguro. 2004. Proclitic contexts in European Portuguese and their effect on clitic placement. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*, 334–352. Stanford, CA: CSLI Publications.
- Luís, Ana & Ryo Otaguro. 2005. Morphological and syntactic well-formedness: The case of European Portuguese clitics. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '05 conference*, 253–270. Stanford, CA: CSLI Publications.
- Luís, Ana & Louisa Sadler. 2003. Object clitics and marked morphology. In Claire Beyssade, Olivier Bonami, Patricia Cabredo Hofherr & Francis Corblin (eds.), *Empirical issues in syntax and semantics*, vol. 4, 133–153. Paris: Presses Universitaires de Paris-Sorbonne.
- Luís, Ana & Andrew Spencer. 2005. A paradigm function account of ‘mesoclisism’ in European Portuguese. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 2004*, 177–228. Dordrecht: Kluwer Academic Publishers. DOI: 10.1007/1-4020-2900-4\_7.
- Maling, Joan & Annie Zaenen (eds.). 1990. *Modern Icelandic syntax*. Vol. 24 (Syntax and Semantics). San Diego, CA: Academic Press. DOI: 10.1163/9789004373235.
- Matsumoto, Yo. 1992. *On the wordhood of complex predicates in Japanese*. Stanford, CA: Stanford University. (Doctoral dissertation).
- Mayer, Elisabeth. 2006. Optional direct object clitic doubling in Limeño Spanish. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '06 conference*. Stanford, CA: CSLI Publications.
- Miller, Philip H. 1992. *Clitics and constituents in phrase structure grammar*. New York: Garland.
- Miller, Philip H. & Ivan A. Sag. 1997. French clitic movement without clitics or movement. *Natural Language & Linguistic Theory* 15(3). 573–639.
- Mohanan, Tara. 1990. *Arguments in Hindi*. Stanford, CA: Stanford University. (Doctoral dissertation).
- Mohanan, Tara. 1994. *Argument structure in Hindi* (Dissertations in Linguistics). Stanford, CA: CSLI Publications.

Alex Alsina

- Monachesi, Paola. 1999. *A lexical approach to Italian cliticization*. Stanford, CA: CSLI Publications.
- O'Neill, Paul. 2016. Lexicalism, the principle of morphology-free syntax and the principle of syntax-free morphology. In Andrew Hippisley & Gregory Stump (eds.), *The Cambridge handbook of morphology* (Cambridge handbooks in language and linguistics), 237–271. Cambridge, UK: Cambridge University Press.
- Patejuk, Agnieszka & Adam Przepiórkowski. 2016. Reducing grammatical functions in LFG. In Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King & Stefan Müller (eds.), *Proceedings of the joint 2016 conference on Head-Driven Phrase Structure Grammar and Lexical Functional Grammar*, 541–559. Stanford, CA: CSLI Publications.
- Perlmutter, David M. 1978. Impersonal passives and the unaccusative hypothesis. In Jeri J. Jaeger, Anthony C. Woodbury, Farrell Ackerman, et al. (eds.), *Proceedings of the 4th annual meeting of the Berkeley Linguistics Society*, 157–189. Berkeley: University of California. DOI: 10.3765/bls.v4i0.2198.
- Perlmutter, David M. 1983. Personal vs. impersonal constructions. *Natural Language & Linguistic Theory* 1. 141–200. DOI: 10.1007/bf00210379.
- Perlmutter, David M. 1989. Multiattachment and the unaccusative hypothesis: The perfect auxiliary in Italian. *Probus* 1. 63–119. DOI: 10.1515/prbs.1989.1.1.63.
- Poletto, Cecilia & Christina Tortora. 2016. Subject clitics. In Adam Ledgeway & Martin Maiden (eds.), *The Oxford guide to the Romance languages*, 772–785. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199677108.003.0047.
- Quaglia, Stefano. 2012. On the syntax of some apparent spatial particles in Italian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 503–523. Stanford, CA: CSLI Publications.
- Renzi, Lorenzo & Laura Vanelli. 1983. I pronomi soggetto in alcune varietà romanze. In Paola Benincà, Manlio Cortelazzo, Aldo Prosdocimi, Laura Vanelli & Alberto Zamboni (eds.), *Scritti linguistici in onore di Giovan Battista Pellegrini*, 121–145. Pisa: Pacini.
- Rizzi, Luigi. 1976. Ristrutturazione. *Rivista di Grammatica Generativa* 1. 1–54.
- Rizzi, Luigi. 1982. *Issues in Italian syntax*. Dordrecht: Foris Publications. DOI: 10.1515/9783110883718.
- Rizzi, Luigi. 1997. The fine structure of the left periphery. In Liliane Haegeman (ed.), *Elements of grammar: Handbook in generative syntax*, 281–337. Dordrecht: Kluwer Academic Publishers. DOI: 10.1007/978-94-011-5420-8\_7.
- Rosen, Carol G. 1984. The interface between semantic roles and initial grammatical relations. In David M. Perlmutter & Carol G. Rosen (eds.), *Relational grammar*, vol. 2, 38–77. Chicago: University of Chicago Press.



- Saccon, Graziella. 1995. *Ne-cliticization does not support the unaccusative/intransitive split*. In Glyn Morrill & Richard Oehrle (eds.), *Formal grammar: Proceedings of the conference of the European Summer School in Logic, Language, and Information*, 227–238. Barcelona: Universitat Politècnica de Catalunya.
- Schwarze, Christoph. 1996. The syntax of Romance auxiliaries. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '96 conference*. Stanford, CA: CSLI Publications.
- Schwarze, Christoph. 2001a. Do sentences have tense? In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '01 conference*, 449–463. Stanford, CA: CSLI Publications.
- Schwarze, Christoph. 2001b. On the representation of French and Italian clitics. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '01 conference*, 280–304. Stanford, CA: CSLI Publications.
- Sells, Peter. 2013. Lexical-Functional Grammar. In Marcel den Dikken (ed.), *The Cambridge handbook of generative syntax*, 162–201. Cambridge, UK: Cambridge University Press.
- Todolí, Júlia. 2002. Els pronoms. In Joan Solà, Maria Rosa Lloret, Joan Mascaró & Manuel Pérez Saldanya (eds.), *Gramàtica del català contemporani*, vol. 2, 1337–1433. Barcelona: Editorial Empúries.
- Tuten, Donald N., Enrique Pato & Ora R. Schwarzwald. 2016. Spanish, Astur-Leonese, Navarro-Aragonese, Judaeo-Spanish. In Adam Ledgeway & Martin Maiden (eds.), *The Oxford guide to the Romance languages*, 382–410. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199677108.003.0022.
- Vallduví, Enric. 2002. L'oració com a unitat informativa. In Joan Solà, Maria Rosa Lloret, Joan Mascaró & Manuel Pérez Saldanya (eds.), *Gramàtica del català contemporani*, vol. 2, 1221–1279. Barcelona: Editorial Empúries.
- Vanhoe, Henk. 2002. Aspects of the syntax of psychological verbs in Spanish: A lexical functional analysis. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '02 conference*, 373–389. Stanford, CA: CSLI Publications.
- Vincent, Nigel. 2019. CP and COMP in diachrony. In Miriam Butt, Tracy Holloway King & Ida Toivonen (eds.), *Proceedings of the LFG '19 conference*, 314–333. Stanford, CA: CSLI Publications.
- Wescoat, Michael T. 2007. Preposition-determiner contractions: An analysis in Optimality-Theoretic Lexical-Functional Grammar with lexical sharing. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '07 conference*, 439–459. Stanford, CA: CSLI Publications.
- Yang, Fengrong. 2019. *Argument realization: Grammatical function and case assignment*. Barcelona: Pompeu Fabra University. (Doctoral dissertation).

*Alex Alsina*

- Yates, Nicholas. 2002. French causatives: A biclausal account in LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '02 conference*, 390–407. Stanford, CA: CSLI Publications.
- Zaenen, Annie, Joan Maling & Höskuldur Thráinsson. 1985. Case and grammatical functions: The Icelandic passive. *Natural Language & Linguistic Theory* 3(4). 441–483. DOI: [10.1007/bf00133285](https://doi.org/10.1007/bf00133285). Reprinted in Maling & Zaenen (1990: 95–136).
- Zipf, Jessica & Stefano Quaglia. 2017. Asymmetries in Italian matrix wh-questions: Word order and information structure. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '17 conference*, 387–405. Stanford, CA: CSLI Publications.
- Zwicky, Arnold. 1987. Slashes in the passive. *Linguistics* 25. 639–665. DOI: [10.1515/ling.1987.25.4.639](https://doi.org/10.1515/ling.1987.25.4.639).



## Chapter 8

# LFG and Semitic languages

Louisa Sadler

University of Essex

This chapter surveys the work in LFG on the Semitic languages of Arabic, Hebrew and Maltese. The overview is structured around a number of themes and topics where there is LFG work on one or more of the Semitic languages. Successive sections look at basic clause structure, verbal complementation (including temporal and aspectual auxiliaries, phasal verbs, and perceptual report verbs), copula constructions, construct state nominals, mixed categories, negation and unbounded dependency constructions.

### 1 Introduction

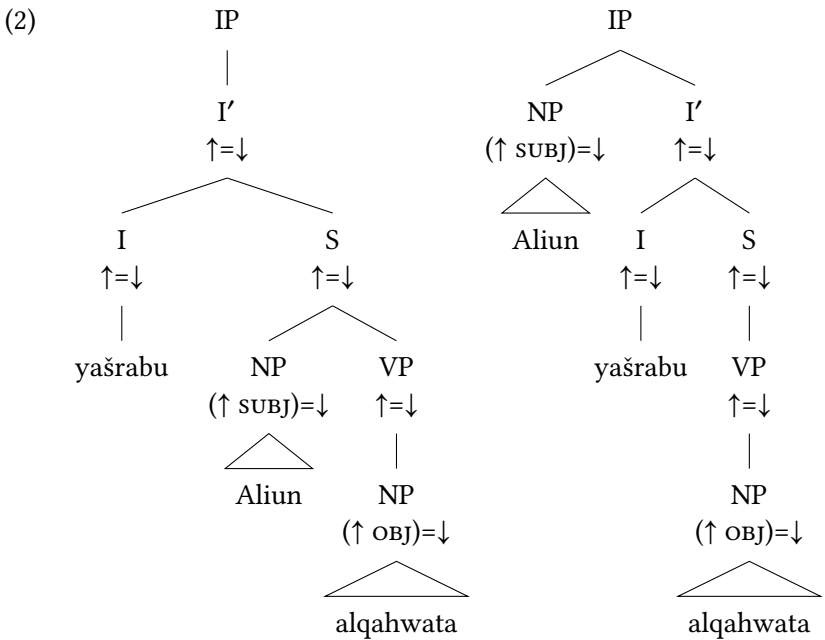
The Semitic languages are part of the Afro-Asiatic family and the genus includes Arabic, Amharic, Tigrinya, Hebrew, Tigré, Maltese, Mehri and Jibbali *inter alia*. Of these, Arabic (including its many modern vernaculars, and the codified, formal variety Modern Standard Arabic (MSA)) is spoken over a very extensive geographical area with in the order of 250–300 million native language speakers, while Amharic, Tigrinya, Hebrew and Tigré all have numbers of speakers in excess of 1 million. Most work in LFG on this family is on (Modern) Hebrew, Arabic (Modern Standard (MSA) and the modern vernaculars) and Maltese (a mixed language with a Maghrebi/Siculo-Arabic stratum). Kifle (2007) and Kifle (2011) are concerned respectively with differential object marking and the applicative construction in Tigrinya, a Semitic language of Eritrea and Ethiopia; see **chapters/African** for further discussion of Tigrinya.<sup>1</sup>

---

<sup>1</sup>Example sentences in this chapter have been taken from a number of different sources. In each case, the examples are given using the author's own transcription, with the exception of long vowels, where the notation has been standardised. On the other hand, some standardisation of glossing has been adopted to increase transparency.





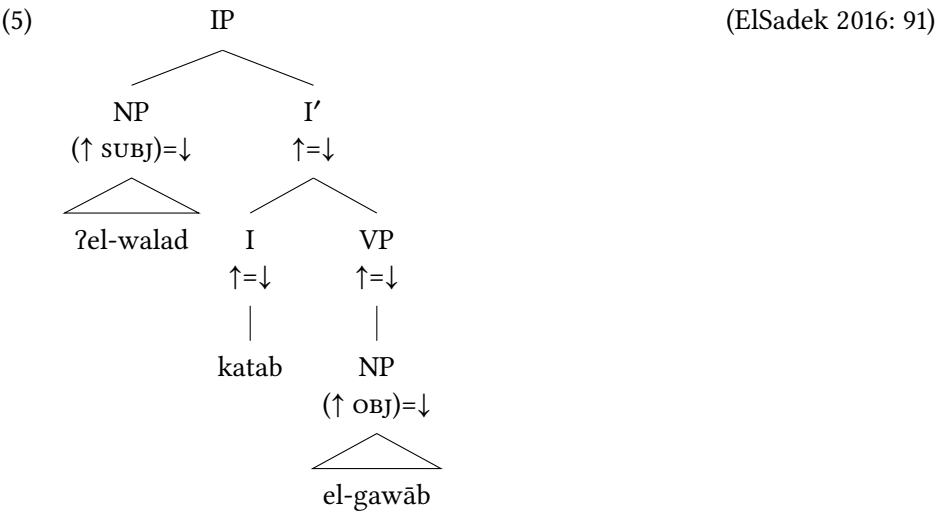


In a slight variant, both ElSadek (2016) and Alruwaili (2019) assume an I+VP structure for the basic neutral svo word order in Egyptian Arabic and Turaif Arabic respectively.<sup>3</sup>

- (3) Egyptian Arabic (ElSadek 2016: 90)  
 ?el-walad katab el-gawāb  
 DEF-boy write.PFV.3SGM DEF-letter  
 'The boy wrote the letter.'
- (4) Turaif Arabic (Alruwaili 2019: 100)  
 ʕali kiteb l-wāḡib  
 Ali write.PFV.3SGM DEF-homework  
 'Ali wrote the homework.'

<sup>3</sup>As with many other vernaculars, vso is a possible but less common variant in Turaif Arabic.

Louisa Sadler



In all of this work, an important motivation for the assumption that the verb expressing tense is in I is the fact that the very same (perfective and imperfective) forms express aspectual information when they occur in a lower position, in the compound tenses of Arabic (the examples (1a) and (6) provide a simple illustration of this property). There is some discussion of compound tenses in Arabic (involving forms of the ‘be’ verb as a temporal auxiliary) in a number of LFG sources and this literature includes both Aux-feature and Aux-PRED analyses for broadly comparable data across the dialects.

Alsharif (2014) adopts a single-tier or Aux-feature analysis for MSA examples such as (6), and a fuller development of this approach to compound tense formation in MSA is given in Alsharif & Sadler (2009).<sup>4</sup>

- (6) MSA (Alsharif 2014: 52)
- |             |         |                        |                |
|-------------|---------|------------------------|----------------|
| kāna        | Ali-un  | ya-šrab-u              | al-qahwat-a    |
| be.PFV.3SGM | Ali-NOM | 3M-drink.IPFV-SG.INDIC | DEF-coffee-ACC |
- ‘Ali was drinking the coffee.’

- (7) MSA (Alsharif & Sadler 2009: 18)

<sup>4</sup>In the simple tenses of Arabic, the imperfective and perfective forms of the lexical verb are associated with TENSE. The compound tenses of Arabic and Maltese are formed by combining imperfective and perfective verb forms of the auxiliary ‘be’ (associated with TENSE) with perfective and imperfective forms of the lexical verb, which are then associated with ASPECT. Note that these forms still show subject agreement in their (embedded) aspectual use.

## 8 LFG and Semitic languages

kun-tu      ʔaktub-u      t-taqrīr-a  
 be.PFV-1SG write-IPFV.1SG the-report-ACC  
 ‘I was writing the report.’

- (8) 
$$\left[ \begin{array}{ll} \text{PRED} & \text{'WRITE<SUBJ,OBJ>'} \\ \text{ASP} & \text{PROG} \\ \text{TENSE} & \left[ \text{PAST} \text{ +} \right] \\ \text{SUBJ} & \left[ \begin{array}{ll} \text{PERS} & 1 \\ \text{NUM} & \text{SG} \end{array} \right] \end{array} \right]$$
 (Alsharif & Sadler 2009: 18)

The Aux-feature account is also adopted by Alotaibi (2014) for Hijazi (Taif) Arabic and Alruwaili (2019) for Turaif Arabic, and by Camilleri (2016) for Maltese. In (10) the auxiliary elements *kont* ‘be.PFV.1SG’ and *qed* respectively contribute TENSE=PAST and ASPECT=PROG to the f-structure of the predicate *wash*.

- (9) Hijazi (Taif) Arabic (Alotaibi 2014: 37)  
 ʔahmad kân      yġri      fī al-hadiqah ʔams  
 Ahmad be.PFV.3SGM run-IPFV.3SGM in DEF-garden yesterday  
 ‘Ahmad was running in the garden yesterday.’

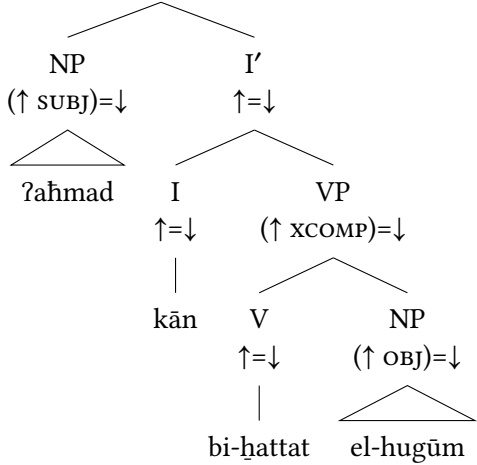
- (10) Maltese (Camilleri 2016: 19)  
 Kon-t      qed      n-a-ħsel      il-karozza  
 be.PFV-1SG PROG 1-FRM.VWL-wash-IPFV.SG DEF-car  
 ‘I was washing the car.’

On the other hand, ElSadek (2016) presents some arguments in favour of the Aux-PRED analysis for Egyptian Arabic, in which the tense-aspect auxiliary *kân* is treated as a raising verb taking a VP xCOMP complement. The c-structure for (11) and f-structure for (13) below illustrate this approach. In work on the aspectual system of Libyan Arabic, Börjars et al. (2016) also provide arguments in support of an Aux-PRED approach to the facts which they discuss.

- (11) Egyptian Arabic (ElSadek 2016: 91)  
 ʔahmad kân      bi-yħattat      el-hogūm  
 Ahmed be.PFV.3SGM BI-plan-IPFV.3SGM DEF-attack  
 ‘Ahmed was planning the attack.’

Louisa Sadler

(12) (ElSadek 2016: 91)



(13) Egyptian Arabic (ElSadek 2016: 90)

konna ḥa-nmût  
be.PFV.1PL FUT-die.IMPV.1PL  
'We were going to die.'

(14) 

PRED	'BE<XCOMP> SUBJ'						
TENSE	PAST						
SUBJ	<table border="1" style="display: inline-table; vertical-align: middle;"><tr><td>PERS</td><td>1</td></tr><tr><td>NUM</td><td>PL</td></tr></table>	PERS	1	NUM	PL		
PERS	1						
NUM	PL						
XCOMP	<table border="1" style="display: inline-table; vertical-align: middle;"><tr><td>PRED</td><td>'DIE&lt;SUBJ&gt;'</td></tr><tr><td>ASPECT</td><td>PROSP</td></tr><tr><td>SUBJ</td><td></td></tr></table>	PRED	'DIE<SUBJ>'	ASPECT	PROSP	SUBJ	
PRED	'DIE<SUBJ>'						
ASPECT	PROSP						
SUBJ							

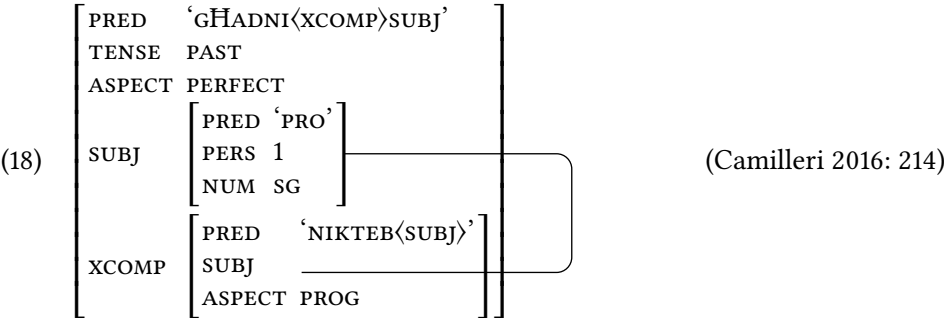
 (ElSadek 2016: 90)

### 3 Aspects of Verbal Complementation

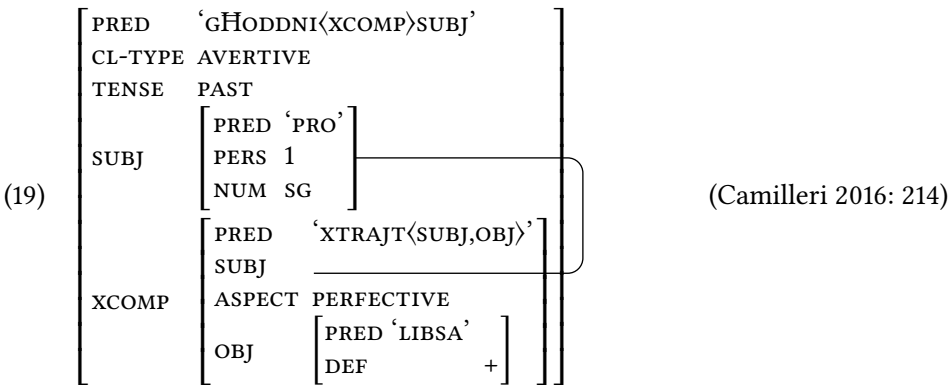
Various further aspects of verbal complementation in the Arabic vernaculars are discussed in the LFG literature. Camilleri (2016) provides a detailed exploration of temporal and aspectual auxiliiation in Maltese, articulating an unusually large set of features and values for this domain at f-structure. She also explores the use of the pseudo-verbs *ghodd-* 'almost' *il-* 'to' and *ghad-* 'still' as aspectual auxiliaries expressing the PERFECT aspect. The term pseudo-verb is used descriptively in work on the Arabic vernaculars to refer to a form which plays the role of a finite verb in the syntax but which is derived from a participle, preposition or nominal stem and usually retains many aspects of morphosyntactic realization reflecting

this origin, such as exhibiting non-canonical forms of subject agreement. These forms raise many interesting issues for analysis, not least regarding their synchronic categorial identity. Camilleri argues that the universal perfect and the perfect of recent past are expressed syntactically in Maltese by the pseudo-verbs *il-* and *ghad-* respectively (see (15) and (16)), while *ghodd* provides an averted construction. Applying a range of standard tests, she argues for an Aux-PRED, raising analysis of these forms, along the lines shown in (18) and (19) for (16) and (17) respectively. Note that Maltese, like Arabic, lacks an infinitival form, and makes use of the imperfective form of the verb in these non-finite complements.

- (15) Maltese (Camilleri 2016: 205)  
Il-ni            n-i-kteb                            mis-7  
to-1SG.ACC 1-FRM.VWL-write.IPFV.SG from.DEF-7  
'I have been writing since 7 o'clock.'
- (16) Maltese (Camilleri 2016: 213)  
Kon-t            għad-ni            qed n-i-kteb  
be.PFV-1SG still-1SG.ACC PROG 1-FRM.VWL-write.IPFV.SG  
'I was still writing.'
- (17) Maltese (Camilleri 2016: 213)  
Kon-t            għodd-ni    xtraj-t            il-libsa  
be.PFV-1SG almost-1SG buy.PFV-1SG DEF-dress  
'I had almost bought the dress.'



Louisa Sadler



The syntax and morphosyntax of phasal verbs, that is verbs which denote the inception, duration, continuation, completion or termination of a state or event (such as (20)), in the Arabic vernaculars is addressed in Alotaibi et al. (2013) (see also Camilleri 2016 and ElSadek 2016 for more extensive discussion of Maltese and Egyptian respectively). These verbs take verbal complements (or, particularly in Modern Standard Arabic, nominalised verbal complements) and typically disallow intervening material between the aspectual verb and its verbal complement (which generally lacks a complementising particle). The aspectual verb and the embedded verb have the same subject, which is not expressed as an NP in the lower clause. The embedded verb shows subject agreement and is usually an imperfective form (Arabic lacks an infinitive form). Using standard tests, Alotaibi et al. (2013) show that a raising analysis is motivated for these verbs in examples such as (20–21a) below.<sup>5</sup>

(20) Egyptian Arabic (Alotaibi et al. 2013: 17)

- a. el-walad ma-bada?-š                      ya-kul  
DEF-boy NEG-start.PFV.3SGM.NEG 3-eat.IPFV.SGM  
'The boy didn't start to eat.'
- b. el-walad bada?                      ma-ya-kul-š  
DEF-boy start.PFV.3SGM NEG-3-eat.IPFV.3SGM.NEG  
'The boy started to not eat.'

(21) a. Hijazi Arabic (Alotaibi et al. 2013: 20)

<sup>5</sup>In addition to occurring in a raising structure, some of the class of phasal verbs also occur in a 'subjectless' variant with a default 3SGM phasal verb and a subject expressed within the embedded complement, a structure which provides an expletive subject counterpart to the raising structure.



## 8 LFG and Semitic languages

al-maḥṣūl bada ya-n-ḡimf  
 DEF-harvest start.PFV.3SM 3-PASS-gather.IPFV.SGM  
 ‘The harvest started being gathered.’

## b. Maltese (Alotaibi et al. 2013: 20)

L-iltiema bde-w j-i-n-ḡabr-u  
 DEF-orphans begin.PFV.3-PL 3-FRM.VWL-PASS-gather.IPFV-PL  
 ‘The orphans started being gathered (together).’

Camilleri et al. (2014b) discuss perceptual report predicates in MSA and in Maltese. The MSA verb *yabdū* ‘seem, appear’ occurs in an expletive subject (or ‘subjectless’) construction taking a complement introduced by the declarative complementising particle *ʔanna*. While it does not permit subject raising (SSR) they argue that it does permit copy raising (CR) with the complementising particle *kaʔanna* ‘as if’. In the CR construction, the copy pronoun is not restricted to the embedded SUBJ role and may occur in a wide range of nominal GFS in the embedded complement.

In Maltese the perceptual report predicates include the verb *deher* ‘seem/appear’ and the pseudo-verbs *donn*+PRN (diachronically the imperative of ‘believe/think’) and *qis*+PRN, both meaning ‘seem/appear/taste/sound as.though/as.if’. (22) exemplifies the expletive construction with the verb *deher*, in which the verb appears in the default 3SGM form and the subject is expressed only in the embedded COMP. In (23) the subject is in the matrix clause and both matrix and embedded verbs agree with it. Camilleri et al. (2014b) argue that evidence from standard tests for raising (idiom chunks, meaning preservation under passivisation, expletives, etc) suggests that (23) and similar examples are SSR.

## (22) Maltese (Camilleri et al. 2014b: 191)

J-i-dher t-tfal sejr-in tajjeb  
 3-FRM.VWL-appear.IPFV.SGM DEF-children going.ACT.PRT-PL good.SGM  
 ‘It seems the children are doing well.’

## (23) Maltese (Camilleri et al. 2014b: 191)

It-tfal dehr-u qed j-iehd-u gost  
 DEF-children appear.PFV.3-PL PROG 3-take.IPFV-PL pleasure  
 ‘The children seem (as though) they are enjoying themselves (lit: taking pleasure).’

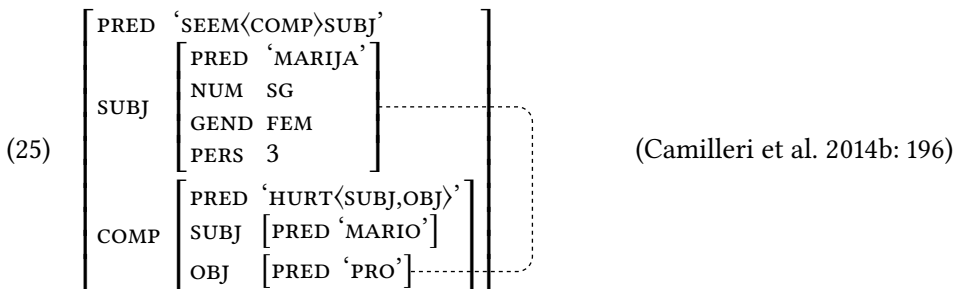
However Maltese *deher* also occurs in what looks like a copy raising (CR) construction, in which a pronominal coreferential with the SUBJ of the raising predi-

*Louisa Sadler*

cate *deher* occurs as an argument within the embedded complement. This is illustrated in (24) where the OBJ pronominal inflection *-ha* in the form *weggagh-ha* ‘hurt.CAUSE.PFV.3SGM-3SGF.ACC’ is coferential with the (inflectionally-expressed) matrix SUBJ (indicated by the dashed line between anaphor and antecedent in (25)).

- (24) Maltese (Camilleri et al. 2014b: 195)

Marija t-i-dher                      wegğagh-ha                      sew, Mario  
Mary 3-FRM.VWL-appears.SGF hurt.CAUSE.PFV.3SGM-3SGF.ACC well Mario  
'Mary seems as though Mario hurt her a lot.'



The analysis which Camilleri et al. (2014b) develop of the syntax and semantics of these perceptual report predicates builds on Asudeh & Toivonen (2012)'s work on English and Swedish. Because Maltese (and Arabic in general) is both a pro-drop language and uses the imperfective form of the verb in non-finite complement clauses (lacking an infinitive form), examples such as (23) could in principle involve either raising or copy raising. They argue that there is a clear contrast between SSR examples such as (23), in which any aspect of the eventuality can be the perceptual source, and CR examples such as (26) in which it is the raised SUBJ itself that is necessarily the individual PSOURCE. In the Maltese CR construction, the pronominal copy can correspond to a very wide range of embedded functions. It is also not limited to the immediately embedded COMP but within the topmost embedded COMP it is restricted to non-subject functions.

- (26) Maltese (Camilleri et al. 2014b: 192)

T-i-dher                      ġa              ta-w-ha                      xebġha  
 3-FRM.VWL-seem.IPFV.SGF already give.PFV.3-PL-3SGF.ACC smacking  
 xoġhol x't-a-ġħmel!  
 work    what.3-FRM.VWL-do.IPFV.SGF  
 'She<sub>i</sub> seems like they already gave her<sub>i</sub> a whole load of work to do!'

## 8 LFG and Semitic languages

While Camilleri et al. (2014b) are concerned with canonical verbal perceptual report predicates in MSA and Maltese, ElSadek & Sadler (2015) look at the expression of perceptual reports in Egyptian Arabic using the active participle *bāyen* ‘show, appear’ and in particular at the use of the (noun-derived) pseudo-verb *fakl* (>‘form, shape’) as a perceptual report predicate. *bāyen* can occur in a construction in which the active participle is followed by a PP which expresses the (visible) individual PSOURCE with either the standard sentential complementiser *?in* (corresponding to the MSA complementiser *?anna*) or the ‘evidential’ complementiser *ka?in* (cognate with MSA *ka?anna*). The active participle must be in the default form but a temporal auxiliary may agree with the nominal PSOURCE in the PP, as illustrated in (27), in what may be a case of parasitic or miscreant agreement.

- (27) Egyptian Arabic (ElSadek & Sadler 2015: 92)  
 konti            bāyen                    ?alē-ki ?inn-ik    mabsūt-a  
 be.PFV.2SGF show.ACT.PCTP.SGM on-2SGF that-2SGF happy.PASS.PTCP.SG-F  
 ‘You seemed happy.’

With *fakl*, there is rather clearer evidence of raising. (28) illustrates a very common means of expressing a perceptual report. It involves what appears morphosyntactically to be a nominal form *šakl* ‘form, shape’ with a dependent ‘possessor’ corresponding to the individual about whom the report is made. Notice in (28) that it is the dependent ‘possessor’ (the pronominal affix) which controls agreement on the ACT.PTCP, and similarly in an example such as (29). Synchronically, this form appears to operate as a pseudo-verb here, in a raising structure.

- (28) Egyptian Arabic (ElSadek & Sadler 2015: 95)  
 fakl-ohom mestaney-în            hāga mohemma  
 form-3PL wait.ACT.PTCP-PL thing important  
 They seem to be waiting for an important thing =  
 ‘It seems they’re waiting for an important thing.’

- (29) Egyptian Arabic (ElSadek & Sadler 2015: 98)  
 fakl el-welād kānu            biyitderbo  
 form DEF-boys be.PFV.3PL beat.BI.IPFV.PASS.3PL  
 ‘The boys seem to have been (being) beaten.’

In structures such as (28) and (29) the dependent NP or pronoun is not obligatorily interpreted as the individual PSOURCE. In a different structure, illustrated in

Louisa Sadler

(30), we find a sentential complement introduced by the complementising particle *kaʔin*, with no requirement that the dependent NP/pronoun be co-referential with the subject of the (embedded) predication, and these structures *are* associated with a clear individual PSOURCE interpretation.

- (30) Egyptian Arabic (ElSadek & Sadler 2015: 98)  
 ʃakl el-welād kaʔenn-aha darabet-hom  
 form DEF-boys as.if-3SGF beat.PFV.3SGF-3PL  
 ‘The boys seem as if she’s beaten them.’

Other work on aspects of complementation includes the following. ElSadek (2016) discusses the causative *χalla* ‘make’, aspectual/phasal verbs and modal verbs, proposing analyses involving functional and anaphoric control. Alotaibi et al. (2013) concerns the description and analysis of experiencer-object psychological predicates (*frighten* or *please* class – EOPVs) in Hijazi Arabic, Egyptian Arabic and Maltese and proposes that the interaction of EOPVs with aspectual raising predicates involves copy raising (CR). An analysis of aspectual object marking in Libyan Arabic is provided in Börjars et al. (2016). In Libyan Arabic, the presence of the preposition *fi* before the direct object of a transitive verb in the imperfective form provides a continuous or habitual aspectual value to the clause (see (31)), which Börjars et al. (2016) model by means of a clause feature INTERIOR=+.

- (31) Libyan Arabic (Börjars et al. 2016: 126)
- |    |                                  |               |                 |
|----|----------------------------------|---------------|-----------------|
| a. | aḥmed                            | kle           | el-koski        |
|    | Ahmed                            | eat.PST.3SGM  | DEF-couscous    |
|    | ‘Ahmed ate couscous.’            |               |                 |
| b. | aḥmed                            | yākil         | fi el-koski     |
|    | Ahmed                            | eat.NONT.3SGM | FI DEF-couscous |
|    | ‘Ahmed eats/is eating couscous.’ |               |                 |

## 4 Copula Sentences

Both Hebrew and Arabic have copula sentences without an overt copula head, as well as copula sentences with a ‘pronominal copula’, and a variety of copula-type elements which mark existential constructions of various sorts. Predicative (copula) sentences with no copula receive present tense interpretations, while an appropriate form of *be* signals other temporal interpretations. The examples in (32) illustrate this alternation between the ‘null’ and overt copula in Hebrew with adjectival, nominal and prepositional predicates.

## 8 LFG and Semitic languages

- (32) a. Hebrew (Falk 2004: 227)  
 Pnina nora xamuda/ tinoket/ b-a-bayit  
 Pnina awfully cute.F/ baby.F/ in-DEF-house  
 ‘Pnina is awfully cute/a baby/in the house.’  
 b. Pnina hayta nora xamuda/ tinoket/ b-a-bayit  
 Pnina be.PST.3SGF awfully cute.F/ baby.F in-DEF-house  
 ‘Pnina was awfully cute/a baby/in the house.’

As well as the zero realisation in the predicative clauses in (32), the so-called pronominal copula also occurs with predicative complements in Hebrew, as well as with a definite NP complement in an equative copula construction, in paradigmatic opposition with forms of *be* giving temporal interpretations other than the present.<sup>6</sup>

- (33) a. Hebrew (Falk 2004: 227)  
 Pnina hi nora xamuda/ ha-tinoket  
 Pnina PRON.3SGF awfully cute.F/ DEF-baby.F  
 ‘Pnina is awfully cute/the baby.’  
 b. Pnina hayta nora xamuda/ ha-tinoket  
 Pnina be.PST.3SGF awfully cute.F/ DEF-baby.F  
 ‘Pnina was awfully cute/the baby.’

The pronominal copula forms of Hebrew and Arabic have received considerable analytic attention outside LFG. Within LFG, Falk (2004) develops a mixed category analysis of the pronominal copula *hi* and its inflectional counterparts in Hebrew, taking it to be categorially nominal but functionally verbal. It is argued to have categorially mixed properties in taking ‘verbal’ complements (e.g. accusative objects) and heading a constituent with a clausal distribution, but occurring in an N position.<sup>7</sup> (34) is the lexical entry for the copula use of *hi*; the

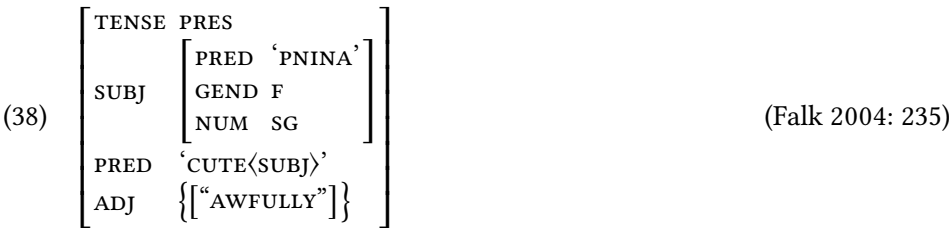
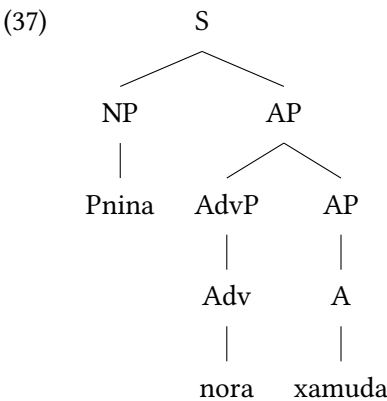
<sup>6</sup>The distribution of the null copula and the pronominal copula strategy in Arabic is similar, but not identical. For example, in Hebrew examples with predicative nominals and PPs are well-formed in the complement of the pronominal copula, but these structures are not found in (most) Arabic vernaculars.

(i) Hebrew (Sichel 1997: 296)  
 Rina hi talmid-a/xaxam-a/b-a-bayit  
 Rina PRON.3SGF student-F/intelligent-F/in-DEF-house  
 ‘Rina is a student/intelligent/at home.’

<sup>7</sup>In Hebrew, the sentential negator *lo* appears before a verb but between the pronominal copula and the following predicative element, which is taken to support the conclusion that the pronominal copula is not a V in c-structure.



of an example such as (32a) which lacks the pronominal copula is along the lines shown in (37–38). On this analysis, non-verbal predication elements which appear in both the null copula and the pronominal copula constructions must be associated with two lexical entries, the predication (i.e. SUBJ-subcategorising) PRED value (for the null copula construction) being a lexical extension of the non-predication one (as can be seen by comparing the relevant PRED values in (36) and (38) respectively).



An interesting consequence of this analysis is that the distinction between individual level predication and stage-level predication is reflected in f-structure. Individual level predication uses the pronominal copula and therefore is associated with a two-tier analysis while stage-level predication (with no copula) is associated with a single simple f-structure (Falk 2004: 236). This contrast in interpretation is illustrated in (39).

- (39) Hebrew (Falk 2004: 236–237)
- a. ha-dinozaur hu vsikor  
DEF-dinosaur PRON.SGM drunk.SGM  
'The dinosaur is a drunkard.'

Louisa Sadler

- b. ha-dinozaur vsikor  
DEF-dinosaur drunk.SGM  
'The dinosaur is drunk.'

Copula clauses with forms of the verb *haya* 'be' are functionally equivalent to both the zero and the pronominal copula constructions, as shown in (32b) and (33b) above. This means that the lexical entry for *haya* must have an optional PRED value (see (40)). As a consequence, a sentence such as (41) will be associated with one c-structure and the two f-structure analyses shown in (42) and (43), that is, it will be analysed as functional ambiguous.

- (40) *hayta*            N    ((↑ PRED) = 'BE<SUBJ, PREDLINK>  
                                  (↑ TENSE) = PAST  
                                  (↑ SUBJ GEND) = F  
                                  (↑ SUBJ NUM) = SG

- (41) Hebrew (Falk 2004: 227)  
Pnina hayta            nora    xamuda  
Pnina be.PST.3SGF awfully cute.F  
'Pnina was awfully cute.'

- (42) 
$$\left[ \begin{array}{ll} \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \left[ \begin{array}{ll} \text{PRED} & \text{'PNINA'} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{SG} \end{array} \right] \\ \text{PRED} & \text{'CUTE<SUBJ>'} \\ \text{ADJ} & \{ \text{["AWFULLY"]} \} \end{array} \right]$$
 (Falk 2004: 237)

- (43) 
$$\left[ \begin{array}{ll} \text{PRED} & \text{'BE<SUBJ, PREDLINK>'} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \left[ \begin{array}{ll} \text{PRED} & \text{'PNINA'} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{SG} \end{array} \right] \\ \text{PREDLINK} & \left[ \begin{array}{ll} \text{PRED} & \text{'CUTE'} \\ \text{ADJ} & \{ \text{["AWFULLY"]} \} \end{array} \right] \end{array} \right]$$
 (Falk 2004: 237)

For MSA, Attia (2008) discusses predicative and locational copula clauses lacking an overt copula form and associates a *be* PRED with the absence of a copula, treating the predicative complement as a PREDLINK. His contention is that the adjective cannot be the head because the subject and the adjective both take what



## 8 LFG and Semitic languages

is considered to be default nominative case, while in the presence of an overt copula the adjective will have accusative case. (44–45) shows this contrast.

- (44) MSA (Attia 2008: 94)  
 al-marʔat-u                      karīmat-un  
 DEF-woman.SGF-NOM generous.SGF-NOM  
 ‘The woman is generous.’
- (45) MSA (Attia 2008: 100)  
 kāna ar-raḡul-u                      karīm-an  
 was DEF-man.SGM-NOM generous.SGM-ACC  
 ‘The man was generous.’

While agreement between the adjective and the clausal subject could be captured simply and transparently by a local SUBJ agreement statement on a two-tier analysis with an open predication complement (that is, an xCOMP analysis along the lines of a raising predicate) this mechanism is not available on the (closed complement) PREDLINK analysis, since the PREDLINK does not contain a SUBJ. Attia (2008) suggests that agreement specifications should be associated with the c-structure rules, as in (46), adapted from Attia (2008: 104).

- (46)
- $$S \rightarrow NP \left\{ \begin{array}{l} \text{VCop} \\ (\uparrow \text{SUBJ}) = \downarrow \end{array} \right. \left| \begin{array}{l} \epsilon \\ (\uparrow \text{PRED}) = \text{'null-be<SUBJ, PREDLINK>'} \\ (\uparrow \text{TENSE}) = \text{PRES} \end{array} \right\} \left\{ \begin{array}{l} \{ NP \mid AP \} \\ (\uparrow \text{PREDLINK}) = \downarrow \\ (\downarrow \text{GEN}) = (\uparrow \text{SUBJ GEN}) \\ (\downarrow \text{NUM}) = (\uparrow \text{SUBJ NUM}) \end{array} \right.$$

The f-structure of a simple predicative copula sentence such as (47) is (48) on this analysis.

- (47) MSA (Attia 2008: 107)  
 huwa ṭālib-un  
 he student.NOM  
 ‘He is a student.’
- (48) 
$$\left[ \begin{array}{ll} \text{PRED} & \text{'NULL-BE<SUBJ, PREDLINK>'} \\ \text{TENSE} & \text{PRES} \\ \text{SUBJ} & [\text{PRED 'HE'}] \\ \text{PREDLINK} & [\text{PRED 'STUDENT'}] \end{array} \right] \quad (\text{Attia 2008: 107})$$

The ‘null-be<SUBJ, PREDLINK>’ analysis is not adopted across the board for the Arabic copula clause. Alsharif (2014) treats verbless predication in MSA with a

Louisa Sadler

single-tier analysis and no ‘null-be’ PRED, as does Alruwaili (2019) for Turaif Arabic. In these analyses the lack of an overt verb is associated simply with TENSE=PRES. Alruwaili (2019) treats the Arabic pronominal copula of equational sentences, illustrated in (49), as an element in I with the PRED value ‘hi<SUBJ,OBJ>’, though without providing much discussion of this analytic choice.

- (49) Turaif Arabic (Alruwaili 2019: 109)  
 huda hī l-mudīr-a  
 Huda COP.3SGF DEF-director-SGF  
 ‘Huda is the director.’

## 5 Construct State Nominals

A considerable theoretical literature addresses the syntax of the *construct state nominal* (or *construct*) (CSN) in Modern Hebrew and Arabic, a construction of central importance in the grammar of these languages. This construction, illustrated in (50)–(52), has a range of distinctive properties: it is left-headed, the head cannot be inflected for definiteness and may occur in a bound form, the *construct state*, depending on language and inflectional class. In MSA the dependent is genitive. A further key property is lack of interruptibility of the head-dependent construction, so that any adjectival modifiers of the head noun follow the entire construct (including any modifiers of the non-head dependent itself), as in example (53). A range of different relations may hold between the head and the non-head or dependent, including possession, partitivity, kinship, identity, measurement and composition, though the range of the construction differs between languages and dialects.<sup>9</sup>

- (50) Hebrew (Falk 2007: 106)  
 mamlexet norvegia  
 kingdom.CONSTR Norway  
 ‘the kingdom of Norway’

---

<sup>9</sup>There are also modificational constructs which get a kind reading as in (i). These are not discussed in any detail in the LFG literature.

- (i) Lebanese Arabic (Ouwayda 2012: 77)  
 abbouʃet sherti  
 hat cop  
 ‘a cop’s type of hat’

8 *LFG and Semitic languages*

- (51) Lebanese Arabic (Ouwayda 2012: 77)  
 sayyaret l-estez  
 car.SGF.CONSTR DEF-teacher  
 ‘the teacher’s car’
- (52) Syrian Arabic (Hallman 2018: 258)  
 ʕamm l-ʕrāus  
 uncle DEF-bride  
 ‘the uncle of the bride’
- (53) Hebrew (Falk 2007: 106)  
 dodat ha-balšan ha-generativi ha-zkena  
 aunt.CONSTR DEF-linguist DEF-generative.M DEF-old.F  
 ‘the generative linguist’s old aunt’
- (54) Jordanian Arabic (Alhailawani 2018: 152)  
 bait il-mara il-jdīd  
 mouse.SGM DEF-woman.SGF DEF-new.SGM  
 ‘the woman’s new house’

As well as the CSN, Hebrew and the Arabic vernaculars have an analytic or free state genitive construction with a distribution which partially overlaps that of the CSN. The following examples illustrate (note that a variety of different “linking elements” are found in the various Arabic vernaculars).

- (55) Lebanese Arabic (Ouwayda 2012: 77)  
 l-sayyara taba? l-estez  
 DEF-car of DEF-teacher  
 ‘the teacher’s car’
- (56) Hebrew (Falk 2007: 104)  
 ha-doda ha-zkena šel ha-balšan  
 DEF-aunt DEF-old of DEF-linguist  
 ‘the old aunt of the linguist’

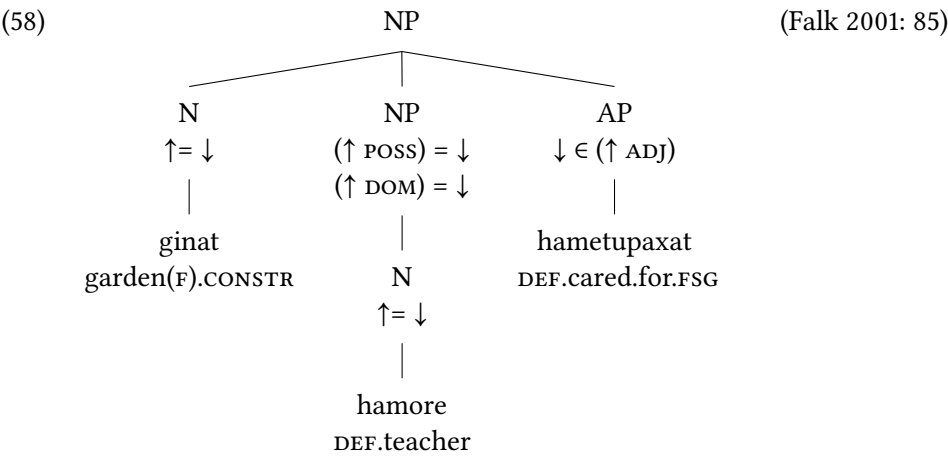
Falk (2001) provides a detailed examination of the constituent structure of NPs containing a *construct* in Hebrew, concluding that despite the closely bound nature of the CSN<sup>10</sup> the N+poss/dependent does not form a constituent to the exclu-

<sup>10</sup>The construct state (of the head noun) is a morphophonological form limited to occurrence within this construction, and within compounds.

Louisa Sadler

sion of the head-modifying AP; the c-structure proposed for (57) is thus (58).<sup>11</sup> The c-structure rule is shown in (59): the  $\downarrow \in (\uparrow \text{ADJ})$  annotation is for the sort of modificational example noted in footnote 9 above which also occur in Hebrew e.g. *bigdey yeladim* ‘clothing.CONSTR children’ (children’s clothing), and is not directly relevant to our discussion below.

- (57) Hebrew (Falk 2001: 85)  
 ginat                    ha-more                    ha-metupax-at  
 garden(F).CONSTR DEF-teacher(M) DEF-cared.for-SGF  
 ‘the teacher’s tended garden’



- (59)  $NP \rightarrow$  N NP AP\* (Falk 2001: 91)  
 $\uparrow = \downarrow$   $(\uparrow \text{DOM}) = \downarrow$   $\downarrow \in (\uparrow \text{ADJ})$   
 $\{ (\uparrow \text{POSS}) = \downarrow \mid$   
 $\downarrow \in (\uparrow \text{ADJ}) \}$

The c-structure rule annotations state that the dependent NP is the value of both a POSS function and a DOM attribute. Nouns are treated as optionally sub-categorising for a POSS, which may be expressed by means of the dependent NP in a CSN, or by means of the alternative free genitive construction. The basic property of the construct form is the tight bond it forms with the dependent (reflected in the choice of a particular variant form of the head noun). Modelling his analysis in part on Wintner (2000)’s use of a DEF attribute in his HPSG analysis, Falk introduces a DOM attribute associated with the immediately post-head

<sup>11</sup>Falk (2007) assumes that any PP modifiers or arguments of the head N are adjoined to the NP, citing a similar proposal developed for Welsh NP structure in Sadler (2000).

constituent. The dependency between the head in the construct state and the dependent NP is thus captured in the f-structure – the construct form (and only this form) selects a *DOM* attribute, which is also the value of the *POSS* feature (the f-description ( $\uparrow\text{DOM}$ ) is an existential constraint, requiring the presence of a *DOM* attribute in the satisfying f-structure). Construct forms cannot occur in other syntactic environments. In a *CSN* the definiteness value of the construction as a whole is “inherited” from the dependent nominal. This is captured in the lexical entry shown in (60) for the construct form of the noun *gina* ‘garden’, i.e. *ginat* by the f-description ( $\uparrow\text{DEF}$ )=( $\uparrow\text{DOM DEF}$ ). The f-structure is shown in (61). In contrast to nouns in construct form, free form nouns are specified as  $\neg(\uparrow\text{DOM})$ .

- (60)

*ginat*

$(\uparrow \text{ PRED}) = \text{'GARDEN'}\langle(\text{POSS})\rangle'$   
 $(\uparrow \text{ NUM}) = \text{SG}$   
 $(\uparrow \text{ GEND}) = \text{F}$   
 $(\uparrow \text{ DOM})$   
 $(\uparrow \text{ DEF}) = (\uparrow \text{ DOM DEF})$

(Falk 2001: 92)

- (61)

[	PRED	'GARDEN'⟨(POSS)⟩'		
	GEND	F		
	NUM	SG		
	DEF	+		
	POSS	CASE	POSS	]
		PRED	'TEACHER'	
		DEF	+	
		GEND	M	
		NUM	SG	
	DOM	}		
ADJ	{[PRED 'OLD']}			

(Falk 2001: 92)

Adjectival modifiers in Hebrew and Arabic show definiteness agreement, in addition to agreement in more canonical agreement features such as *NUM* and *GEND*. In a *CSN* the definiteness value of the construction as a whole is determined by that of the *POSS* or dependent NP, as illustrated in (53), (54) and (57) above. Definiteness agreement is simply captured by associating the relevant inside-out statement (e.g.  $((\text{ADJ } \uparrow) \text{ DEF}=+)$  with the attributive adjective.

Simply put, the essence of Falk (2001)’s analysis is a lexical distinction between construct forms of nouns, which are specified as  $(\uparrow\text{DOM})$  and free forms, which are  $\neg(\uparrow\text{DOM})$  by default, a special *ps* rule which takes care of the adjacency requirement, and the association of the dependent NP with the *POSS* function. Notice that the occurrence of a *POSS* function and the use of the construct form are

*Louisa Sadler*

not co-extensive: some dependent NPs are ADJ, rather than POSS functions, as noted above, and some POSS functions are realised by means of the free genitive construction illustrated in (56) above. It is for this reason that Falk's account separates the requirement for a dependent (DOM) from the function of the dependent (normally POSS).

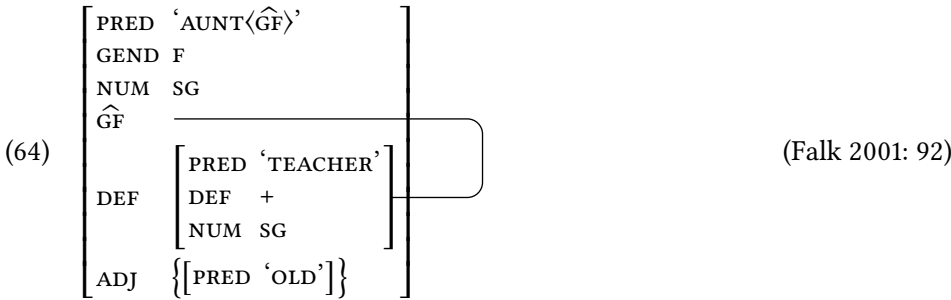
Falk (2007) further develops the analysis of the CSN presented in Falk (2001), providing more extensive discussion of the distribution of the 'short' (i.e. CSN-internal) and 'long' (i.e. *šel*-PP) possessor constructions (i.e. examples such as (56) above). For example, while both constructions are available for relational nouns, true possession in Hebrew is normally expressed by using the *šel* construction (use of the CSN being limited to more formal registers). By contrast, for naming places and periods of time, Hebrew uses only the short construction (see (50)). There are two main theoretical developments, concerning the identification of grammatical functions and the treatment of definiteness and definiteness inheritance.

While Falk (2001) calls the grammatical function of the dependent NP *poss*, Falk (2007) offers a more articulated account, replacing this function by  $\widehat{\text{GF}}$ . The notation  $\widehat{\text{GF}}$  stands for the most prominent argument in an *f*-structure (typically the *SUBJ* in a clausal *f*-structure); Falk (2006) introduces this notation, arguing that the grammatical function *SUBJ* should be deconstructed into the most prominent function, notated  $\widehat{\text{GF}}$  and an ‘overlay’ function, *PIVOT*, a function of cross-clausal connection. The dependent in examples such as (62) involving a relational noun then is treated as the  $\widehat{\text{GF}}$  (rather than *poss*), and the overlay function is argued to be *DEF* (replacing the *DOM* of the earlier account), licensed through structure-sharing (with  $\widehat{\text{GF}}$ ) as stated in (65). As noted above, the head noun in a construct nominal cannot itself be inflected for definiteness and it is the possessor, or  $\widehat{\text{GF}}$  dependent which determines the definiteness of the construction as a whole. (59) is replaced by (63), but expresses essentially the same analysis.<sup>12</sup>

- (62) Hebrew (Falk 2007: 104)  
 dodat            ha-balšan    ha-zkena  
 aunt.CONSTR DEF-linguist DEF-old.F  
 ‘the linguist’s old aunt’

- (63) NP  $\rightarrow$  N NP AP\* Falk (2007: 113)  
 $\uparrow = \downarrow$  ( $\uparrow$  DEF) =  $\downarrow$   $\downarrow \in (\uparrow$  ADJ)  
 $(w(<^*) \text{MORPHTYPE}) = \text{BND}$

<sup>12</sup>The annotation (w (<\*) MORPHTYPE)=BND on the dependent NP specifies that the left sister of the NP's word structure is a bound form.



(65)  $(\uparrow \text{DEF}) = (\uparrow \text{GF}) \mid (\uparrow \text{OBL}_{\text{CON}}) \mid (\uparrow \text{OBL}_{\text{THEME}}) \mid (\uparrow \text{OBL}_{\text{NAME}})$  (Falk 2007: 120)

The re-entrancy stated in (65) takes account of the range of functions which can be expressed within the CSN (replacing the POSS of the previous analysis). An example such as (66) is associated with an  $\text{OBL}_{\text{CON}}$  function (as well as being the value of DEF): other functions which can be expressed by the dependent nominal in a CSN are  $\text{OBL}_{\text{NAME}}$  and  $\text{OBL}_{\text{THEME}}$  – the latter for concrete nouns with a Theme argument as in (67).

- (66) Hebrew (Falk 2007: 117)  
 kos kafe  
 cup coffee  
 ‘a cup of coffee’

- (67) Hebrew (Falk 2007: 122)  
 targumey                      ha-odisea      šel ha-sifriya  
 translation.CONSTR DEF-Odyssey of DEF-library  
 ‘the library’s translation of the Odyssey’

## 6 Mixed Categories

An analysis of the Hebrew action nominal (and NP structure more generally) is offered in Falk (2001) and further developed in Falk (2007). These papers treat action nominals such as (68) as displaying a ‘verbal’ mapping to arguments, signalled by the existence of the ACC-marked OBJ, while others display a purely nominal mapping. In the ‘verbal’ action nominal, the agent argument is realized within the CSN (i.e. as a ‘short’ possessor) or in a *šel*-PP (‘long’ possessor). In each case, it is argued that the c-structure of the action nominal is mixed.

- (68) Hebrew (Falk 2007: 117)

Louisa Sadler

- a. sgirat                    ha-mankal   [et ha-misrad]  
closure.CONSTR DEF-director ACC DET-office  
‘the director’s closure of the office’
- b. ha-sgira      šel ha-mankal   [et ha-misrad]  
DEF-closure of DEF-director ACC DET-office  
‘the director’s closure of the office’

The analysis of an example such as (68a) in Falk (2001) is as follows. The nominal has a mixed c-structure captured in (69), where  $\lambda$  is the category labelling function. A c-structure with both NP and VP projections is required to satisfy this set of constraints, motivating the c-structure rule in (70). Alongside this is the assumption that Hebrew actional nominals have the specification  $(\uparrow \text{ POSS})=(\uparrow \text{ SUBJ})$  and hence the f-structure in (71) arises for the accusative Hebrew actional nominal such as (68a) (given the treatment of dependent NP within the csN developed in Falk 2001). The fundamental insight concerning the f-structure of ‘verbal’ action nominals is that they have a verbal argument structure mapping (e.g. to SUBJ and OBJ) but realise their SUBJ as a POSS.<sup>13</sup> The c-structure proposed by Falk for the ‘verbal’ action nominal is shown in (72).<sup>14</sup>

(69)  $(\uparrow \text{ PRED}) = \text{‘close} \langle \langle x, y \rangle_v \rangle_n$   
 $v: \text{VP} \in \lambda (\phi^{-1} (\uparrow))$   
 $n: \text{NP} \in \lambda (\phi^{-1} (\uparrow))$  (Falk 2001: 96)

(70)  $\text{NP} \longrightarrow \begin{array}{cc} \text{NP} & \text{VP} \\ \uparrow = \downarrow & \uparrow = \downarrow \end{array}$  (Falk 2001: 94)

(71) 

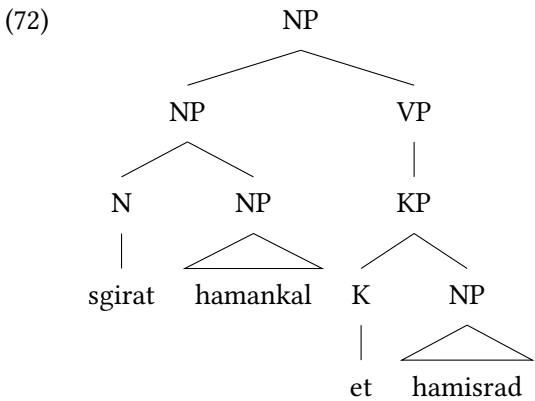
(71)	PRED	‘CLOSE<<X,Y> <sub>v</sub> > <sub>n</sub> ’													
	GEND	F													
	NUM	SG													
	DEF	+													
	DOM	<table border="1" style="display: inline-table; vertical-align: middle;"><tr><td>PRED</td><td colspan="2">‘DIRECTOR’</td></tr><tr><td>DEF</td><td colspan="2">+</td></tr><tr><td>GEND</td><td colspan="2">M</td></tr><tr><td>NUM</td><td colspan="2">SG</td></tr></table>		PRED	‘DIRECTOR’		DEF	+		GEND	M		NUM	SG	
	PRED			‘DIRECTOR’											
	DEF			+											
	GEND	M													
	NUM	SG													
	POSS	<table border="1" style="display: inline-table; vertical-align: middle;"><tr><td colspan="3"> </td></tr></table>													
SUBJ	<table border="1" style="display: inline-table; vertical-align: middle;"><tr><td colspan="3"> </td></tr></table>														
OBJ	[PRED ‘OFFICE’]														

 (Falk 2001: 96)

<sup>13</sup>The argument mapping for (68b) will be similar although there will be no DOM feature because the POSS is not realized within a csN.

<sup>14</sup>As a technical aside, note that although this is a mixed category analysis, according to the standard definition of extended head (Bresnan et al. 2016: 136) the N is not the extended head of the VP, because of the intervening NP node which dominates the csN, a matter which is not discussed in Falk (2001, 2007).





As well as the ‘verbal’ mapping (with an ACC-marked OBJ), Hebrew action nominals may realize their arguments as shown in (73). In (73a) the arg2 or theme is the dependent NP in the construct state nominal, and hence corresponds to a POSS (on the analysis of Falk 2001). This variant has a purely nominal mapping in which the other argument (if present) is an OBL. Hence the PRED value is as shown in (74).

- (73) Hebrew (Falk 2001: 94, 118)
- a. sgirat                      ha-misrad (alyedey ha-mankal)  
     closure.CONSTR DEF-office by              DET-director
  - b. ha-sgira      šel ha-misrad (alyedey ha-mankal)  
     DEF-closure of DEF-office by              DET-director  
     ‘the closure of the office by the director’

(74) (↑PRED) = ‘close<(OBL<sub>AG</sub>), POSS> (Falk 2001: 97)

Evidence that the purely nominal variant also has a mixed c-structure comes from the observation that it can be modified by AdvP as well as by AP, as shown in (75).<sup>15</sup>

- (75) Hebrew (Falk 2001: 98)
- a. ibud              ha-kolot yadanit      alyedey ha-mumxim  
     processing DEF-votes manually by              DEF-experts

<sup>15</sup> Although there is less discussion, Falk (2001) also provides examples showing AP modification of the verbal variant (with the POSS/SUBJ expressed as a *šel* PP), as well as modification by AdvP.

Louisa Sadler

- b. ibud            ha-kolot   ha-yadani   alyedey ha-mumxim  
processing DEF-votes DEF-manual by        DEF-experts  
‘the manual processing of the votes by the experts’

In summary, Falk argues that both ‘verbal’ and ‘nominal’ action nominals in Hebrew have a mixed c-structure. In Falk (2001) the NP realized as the dependent within a CSN nominal (or as a *šel* phrase in the case of ‘long’ possession) is analysed as a POSS, leading to the mappings shown in (76) for the action nominal. Falk (2007) develops a more articulated view of the range of GFs associated with the CSN, as discussed in the previous section, leading to the mappings shown in (77) for the action nominals.

(76)	subcategorisation	additional functions (in CSN)
	lexical description	from the PS rules
verbal mapping	$\langle \text{SUBJ, OBJ} \rangle$ SUBJ = POSS	POSS = DOM
nominal mapping	$\langle \text{OBL}_{AG}, \text{POSS} \rangle$	POSS = DOM

(77)	subcategorisation	additional functions
	lexical description	from the PS rules
verbal mapping	$\langle \widehat{\text{GF}}, \text{OBJ} \rangle$	$\widehat{\text{GF}} = \text{DEF}$
nominal mapping	$\langle \text{OBL}_{AG}, \widehat{\text{GF}} \rangle$	$\widehat{\text{GF}} = \text{DEF}$

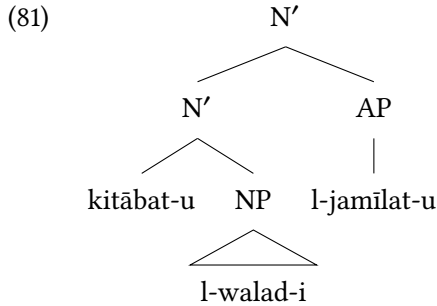
There is relatively little detailed discussion in the LFG literature of the corresponding Arabic NPs, which are headed by *mašdars*. The MSA examples (78) and (79) illustrate the ‘verbal’ and ‘nominal’ mappings respectively.<sup>16</sup>

- (78) MSA (Börjars et al. 2015: 49)  
ʔakl-u        l-walad-i        it-tufāhat-a  
eat.MSD-NOM DEF-boy-GEN DEF-apple-ACC  
‘the boy’s eating the apple’
- (79) MSA (Börjars et al. 2015: 55)  
ʔakl-u        l-walad-i        as-sarīf-u        li-t-tufāhat-i  
eat.MSD-NOM DEF-boy-GEN DEF-fast-NOM of-DEF-apple-GEN  
‘the boy’s fast eating of the apple’

<sup>16</sup>The occurrence of ACC case in (78) is often taken to indicate a mixed categorial status for this construction, with the ‘verbally-marked’ dependent(s) appearing within a VP node.

In connection with his treatment of negation in *maʿṣūl*-headed structures in MSA, Alsharif (2014) adopts Falk (2001)’s analysis of the CSN dependent as a POSS (re-entrant with the DOM feature) and using the additional functional equation POSS=SUBJ for cases in which the head N is a *maṣḍar*, and a mixed category c-structure (at least for the ‘verbal’ *maṣḍar* structures). However he argues for a structure in which the CSN is recognised as a constituent to the exclusion of any adjectival modifiers, as shown in (81) (in contrast to Falk’s (59) above). Börjars et al. (2015) provide agreement data from MSA in support of the same conclusion.

- (80) MSA (Alsharif 2014: 291)  
 kitābat-u        l-walad-i        l-jamīlat-u  
 write.MSD-NOM DEF-boy-GEN DEF-beautiful-NOM  
 ‘the boy’s beautiful writing’



In contrast to the mixed category analysis of Hebrew action nominals developed in Falk (2001, 2007), Börjars et al. (2015) propose a purely nominal c-structure, reflecting the fact that the *maṣḍar* has nominal morphosyntax and may have the external distribution of a NP. The GEN and ACC NPs in the transitive ‘verbal’ *maṣḍar* are both sisters of N – the idea is essentially that of extending the constituent containing the CS to include ACC objects in the case of the ‘verbal’ mapping (all RHS categories are to be interpreted as optional in this rule).<sup>17</sup> The nominal structure in (79) is more hierarchical, with the *li*-PP (corresponding to the second argument of the verb ‘eat’) adjoined at a higher level NP constituent in the structure as an OBL, and the AP also licensed as an ADJunct by a recursive NP → NP XP rule.

<sup>17</sup>Börjars et al. (2015) do not provide an analysis of definiteness inheritance (from the genitive dependent) for the general case of construct state nominals. For the *maʿṣūl*-headed structures of MSA which they are concerned with in this paper they assume the equation (↑DEF) = (↑SUBJ DEF) in the lexical entry of the *maṣḍar*.

Louisa Sadler

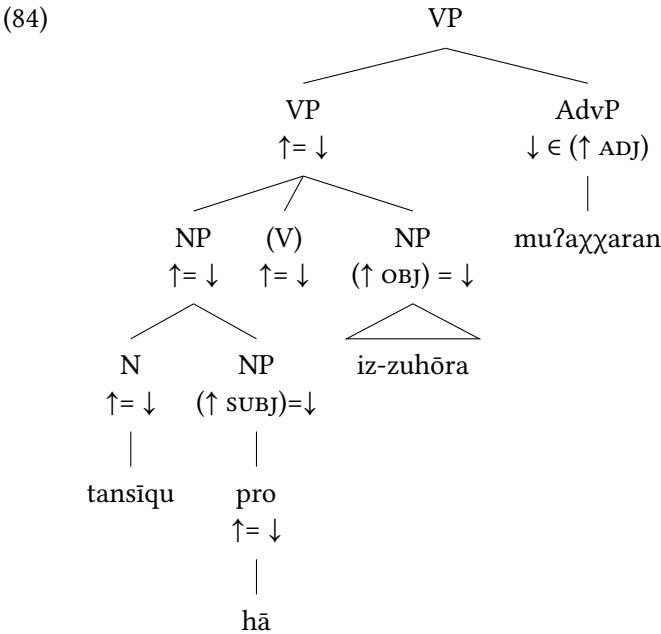
- (82) NP  $\longrightarrow$  N NP NP NP  
 $\uparrow = \downarrow$  ( $\downarrow$ CASE) = GEN ( $\downarrow$ CASE) = ACC ( $\downarrow$ CASE) = ACC  
 $(\uparrow$ SUBJ) =  $\downarrow$  ( $\uparrow$ OBJ) =  $\downarrow$  ( $\uparrow$ OBJ $_{\theta}$ ) =  $\downarrow$   
 (Börjars et al. 2015: 53)

Lowe (2020) points out a number of empirical problems with this analysis, notably in relation to ensuring the correct ordering of any AP and AdvP modifiers in the nominal *maşdar* constructions and in ruling out the occurrence of adjectival modifiers in the ‘verbal’ *maşdar* structures; and also takes issue with it on theoretical grounds. He argues for an approach to mixed category constructions in which internal syntax, rather than morphosyntax or external distribution, is taken to be a sufficient criterion for syntactic categorisation. This leads to a mixed projection (VP over NP) analysis for both types of *maşdar* construction (the VP node is motivated by the presence of an OBJ under the ‘verbal’ mapping and the possibility of adverbial modifiers under both ‘nominal’ and ‘verbal’ mappings). The structures which he proposes, (84) and (86), are rooted in a VP node, despite the nominal nature of the external distribution of these structures.<sup>18</sup>

- (83) MSA (Börjars et al. 2015: 49)  
 tansīq-u =hā iz-zuhōr-a muʔaḫḫaran  
 arrange.MSD-NOM her DEF-flowers-ACC recently  
 ‘her arranging the flowers recently’

<sup>18</sup>To address this issue, Lowe (2020: 333) proposes the use of a complex category  $V_{[msd]}$  and a metacategory in the phrase structure rules to capture the distributional similarity between NPs and *maşdar*-headed VPs. Recall that the meta-category label does not itself give rise to a node in the tree representation, being merely an abbreviatory device.

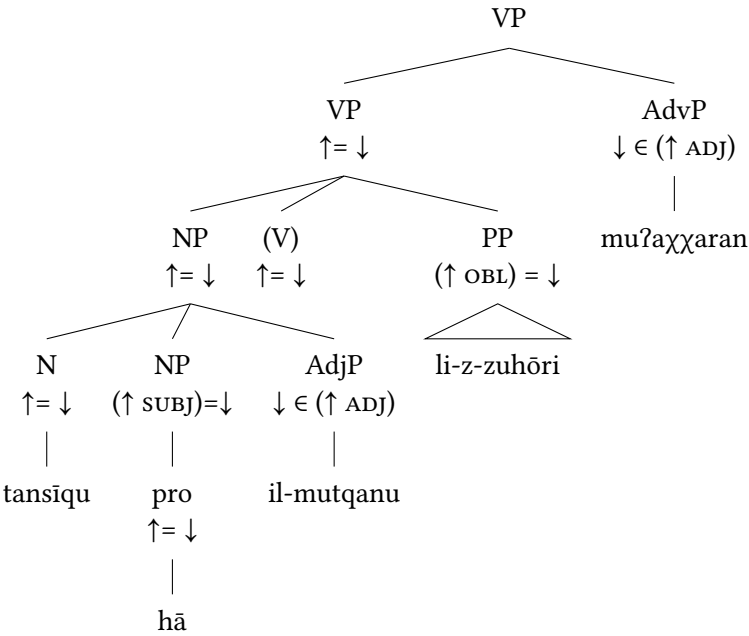
(i)  $NomP \equiv \{NP \mid VP_{[msd]}\}$  (Lowe 2020: 333)



- (85) MSA (Börjars et al. 2015: 55)
- |   |     |                 |                    |            |
|---|-----|-----------------|--------------------|------------|
| tansīq-u  | =hā | il-mutqan-u     | li-z-zuhōr-i       | muʔaxxaran |
| arrange.MSD-NOM                                 | her | DEF-perfect-NOM | of-DEF-flowers-GEN | recently   |
| 'her perfect arranging of the flowers recently' |     |                 |                    |            |

Louisa Sadler

(86)



## 7 Negation

Sentential negation in MSA is expressed by means of the particles *mā*, *lā*, *lan* and *lam* and the inflecting form *laysa* which occurs with both verbal and non-verbal predicates (see (87) and (88)). *laysa* (and its inflectional variants) gives rise to present tense interpretations and shows partial agreement when it precedes the subject and full agreement with a preceding subject, typical verbal behaviour. Accordingly, Alsharif & Sadler (2009) treat *laysa* as a negative (present) tensed verbal element in I.

(87) MSA (Alsharif & Sadler 2009: 10)

- a. al-awlad-u    lays-ū    ya-ktub-ūn  
the-boys-NOM NEG-3MP 3M-write.IPFV-3MP-IND  
'The boys do not write.'
- b. lays-a    al-awlad-u    ya-ktub-ūn  
NEG-3MS the-boys-NOM 3M-write.IPFV-3MP-IND  
'The boys do not write.'

(88) MSA (Benmamoun 2000: 53)

## 8 LFG and Semitic languages

laysa    ʔaḥii        muʔallim-an.  
 NEG.3MS brother.my teacher-ACC  
 ‘My brother is not a teacher.’

The particles *lā*, *lam* and *lan* are strictly verb-adjacent, and do not exhibit agreement with the subject. While *lā* occurs with a verb in the indicative imperfective, *lam* occurs with the jussive imperfective expressing negation in the past, and *lan* with the subjunctive imperfective, expressing negation in the future: thus *lam* and *lan* are negative particles which carry temporal information.

(89) MSA (Benmamoun 2000: 95)

- a. ʔ-ʔullāb-u    laa    ya-drus-uu-n  
          the-students NEG 3M-study.IPFV-3MP-IND  
          ‘The students do not study/are not studying.’
- b. lan        ya-dḥab-a            ʔ-ʔullāb-u  
          NEG.FUT 3M-go.IPFV-MSG.SBJV the-students-NOM  
          ‘The students will not go.’
- c. ʔ-ʔullāb-u        lam        ya-dḥab-uu  
          the-students-NOM NEG.PAST 3M-go.IPFV-MP.JUSS  
          ‘The students did not go.’

Alsharif & Sadler (2009) analyze these negative particles as non-projecting words of category I (notated  $\hat{I}$ ) in the sense of Toivonen (2003), forming a small construction with the immediately following verbal element. The notion of non-projecting word captures the uninterruptibility of the Neg+V sequence, but still treats the negative marker and the verb as separate morphological words. The particles *lam* and *lan* contribute PAST and FUT tense values respectively (and select (tenseless) forms of the verb in a dependent mood), while *lā* cannot co-occur with PAST tense. The negative particle *lan* can also occur as a non-projecting word under V where it contributes not FUT but PROSP aspect. They consider the interaction of these negative particles with both simple and compound tenses in MSA.<sup>19</sup>

<sup>19</sup>A complex TENSE feature with boolean-valued attributes PAST and FUT is adopted in this approach because of the compositional nature of certain periphrastic verb forms. For example, a future tense may be formed periphrastically by combining the imperfective indicative form (which otherwise received a present tense interpretation), with the preverbal particle *sawfa* as in (i), and hence the imperfective indicative is associated with the (underspecified) TENSE PAST=–.

Louisa Sadler

(90) I       $\longrightarrow$        $\hat{I}$       I      Alsharif & Sadler (2009: 14)  
     $\uparrow = \downarrow$        $\uparrow = \downarrow$

(91) *lam*       $\hat{I}$       ( $\uparrow$ TENSE PAST) = +      (Alsharif & Sadler 2009: 16)  
    ( $\uparrow$ POL) = NEG  
    ( $\uparrow$ MOOD) =<sub>c</sub> JUSS

As for MSA *mā*, this marker of sentential negation occurs in sentences with both verbal and non-verbal predicates. It always precedes the predicate but is not required to be immediately adjacent to it. Alsharif (2014) argues that it is a negative complementiser (Arabic has a reasonably extensive range of complementising particles), so that (92) is associated with the c-structure shown in (94).

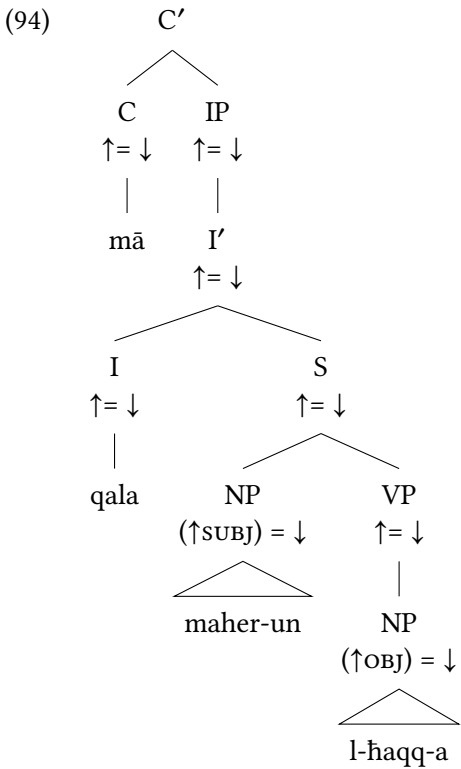
(92) MSA (Alsharif 2014: 169)  
       *mā qal-a maher-un l-ḥaqq-a*  
       NEG say.PFV-3M Maher-NOM DEF-truth-ACC  
       ‘Maher did not say the truth.’

(93) MSA (Alsharif 2014: 132)  
       *mā mohammad-un kātib-un*  
       NEG Mohammad-NOM writer-NOM  
       ‘Mohammad is not a writer.’

---

(i) MSA (Fassi Fehri 1993: 82)  
       *sawfa laa y-ahḍur-u*  
       FUT NEG 3M-present.IPFV-3MS.IND  
       ‘He will not come.’





Adopting the idea that it may mark some sort of contrastive focus as well as negation, (see Ouhalla 1993 and Benmamoun 2000, *inter alia*), Alsharif (2014) also argues that in examples such as (95), the focussed element immediately following the negative complementiser, is in [ Spec,IP ] (in (95) this is the PP *bi-s-sikkīn-i*) (hence this position must host various discourse functions, including that of SUBJ).

- (95) MSA (Alsharif 2014: 173)
- |     |               |           |                |                     |
|-----|---------------|-----------|----------------|---------------------|
| mā  | bi-s-sikkīn-i | jaraḥ-a   | ḫālid-un       | bakr-an             |
| NEG | P-DEF         | knife-GEN | wound.PFV-3SGM | Khalid-NOM Bakr-ACC |
- ‘It is not with a knife that Khalid wounded Bakr.’

The Arabic vernaculars typically use *mā* for negation in verbally-headed sentences, and a set of forms which combine *m-* with pronominal affixes for sentential negation in non-verbal sentences.<sup>20</sup>

<sup>20</sup>The occurrence of verbal negation with many pseudo-verb forms, as in (i), where the literal,

Louisa Sadler

A major split is found across the dialects (roughly between Eastern and Western) according to whether they use a single negative element or bipartite negation, combining an *m*-form with a second marker *-š/-x* which results from grammaticalisation of an earlier form corresponding to *šay?* ‘thing’ in Classical Arabic.

The vernacular verbal negative marker *mā* illustrated in (96) is treated as a non-projecting word in Alsharif (2014) (for Hijazi) and Alruwaili (2019) (for Turaif Arabic), that is, as a syntactic element appearing strictly adjacent to a verbal element.<sup>21</sup>

- (96) Turaif Arabic (Alruwaili 2019: 162)  
ʔali mā kitəb l-wāḡib  
Ali NEG write.PFV.3SGM DEF-homework  
‘Ali did not write the homework.’

- (97) I Turaif Arabic (Alruwaili 2019: 162)  
↑ = ↓  
Neg I  
↑ = ↓ ↑ = ↓  
mā kitəb

Alruwaili (2019) shows that *mā* can occur before either the auxiliary (*kān* ‘be.PFV’) or the lexical verb in compound tenses (and hence can form a small construction with either I or V), and argues in favour of the ternary branching rule (99) as the negator must precede the tense/aspect particle *rāḥ* when they co-occur. As a marker of sentential negation, *mā* specifies ENEG = + (eventuality negation, see Przepiórkowski & Patejuk 2015).

prepositional meaning of *l-* is ‘to’, shows that their reanalysis from their original category into a verbal category is well advanced.

- (i) Turaif Arabic (Alruwaili 2019: 121)  
ʔ-tullāb mā l-hum ʔašam  
DEF-student.PL NEG have-3PLM.GEN discount  
‘The students do not have a discount.’

<sup>21</sup>Clearly, an affixal analysis of the negative markers might be argued to be appropriate for some other dialects.

- (98) Turaif Arabic (Alruwaili 2019: 166)  
 huda mā rāḥ t-sāfar bukra  
 Huda NEG FUT 3SGF-travel.IPFV tomorrow  
 ‘Huda will not travel tomorrow.’

- (99)  $I' \longrightarrow \widehat{\text{Neg}} \quad \hat{I} \quad I$   
 $\uparrow = \downarrow \quad \uparrow = \downarrow \quad \uparrow = \downarrow$

The example in (100) illustrates the marker of sentential negation for non-verbal predicates (and in equational sentences). Both Alsharif (2014) and Alruwaili (2019) treat this marker (and its inflectional variants) as a negative copula (the lexical entry in (101) is from Alruwaili (2019: 170)).

- (100) Turaif Arabic (Alruwaili 2019: 169)  
 huda mū/mahi fi l-bēt  
 Huda NEG.COP/NEG.COP.3SGF in DEF-house  
 ‘Huda is not in the house.’

- (101)  $mū \quad I \quad (\uparrow \text{ENEG}) = + \quad \text{Turaif Arabic (Alruwaili 2019: 170)}$   
 $\text{VP} \notin \text{CAT}(\uparrow)$   
 $(\uparrow \text{TENSE}) = \text{PRES}$

Camilleri & Sadler (2017a) looks at sentential negation in Maltese and the syntactic behaviour of a group of negative sensitive indefinite items (n-words, NSI) in Maltese. In common with many Western dialects of Arabic, Maltese is a language with bipartite negation, as can be seen in the double marking *ma* ..... -*x* in (102). Synchronically, they argue for Maltese that it is *m-/ma* which realizes negation in Maltese, while the -*x* is essentially some sort of NSI. The strategies for sentential negation of clauses with verbal and non-verbal predicates (including the active participle) respectively are shown in (102) and (103) respectively.

- (102) Maltese (Camilleri & Sadler 2017a: 147)  
 Ma qraj-t-x il-ktieb.  
 NEG read.PFV-1SG-NEG DEF-book  
 ‘I didn’t read the book.’

- (103) Maltese (Camilleri & Sadler 2017a: 147)  
 Mhux ~ mhumix sejr-in.  
 NEG.3SGM.NEG ~ NEG.3PL.NEG go.ACT.PTCP-PL  
 ‘They are not going.’

Louisa Sadler

The paper proposes an analysis of the *xejn* ‘nothing’ series of negative indefinites (including *hadd* ‘no one’, *ebda* ‘no(ne)’ and *imkien* ‘nowhere’) which occur in negative sentences. As the examples in (104) show, the negative marker *ma* is required to express sentential negation, irrespective of the linear order of the n-word vis-à-vis the predicate. This behaviour, and the fact that these n-words may provide negative fragment answers, supports the view that Maltese is a strict negative concord language and the classification of these indefinites as simple NCIS. However, although Maltese uses the bi-partite (*ma....-x*) strategy for negation, as shown in (102) above, *-x* is in fact incompatible with these n-words in the same clause, as shown in (105).

(104) Maltese (Camilleri & Sadler 2017a: 150)

- a. Ilbieraħ    hadd    \*(ma) ġie.  
               yesterday no.one NEG    come.PFV.3SGM  
               ‘No one came yesterday.’
- b. Ilbieraħ    \*(ma) ġie                    hadd.  
               yesterday NEG    come.PFV.3SGM no.one  
               ‘No one came yesterday.’

(105) Maltese (Camilleri & Sadler 2017a: 151)

It-tifla    ma    ra-t(\*-x)                    xejn.  
 DEF-girl NEG see.PFV-3SGF-X nothing  
 ‘The girl saw nothing.’

Long-distance licensing of n-words is felicitous in Maltese (depending on the nature of the subordinate clauses), as in (106), and the same incompatibility with the suffix *-x* is observed.<sup>22</sup>

(106) Maltese (Camilleri & Sadler 2017a: 153)

Ma    smaj-t                    [li    qal-u                    [li  
               NEG hear.PFV-1SG COMP say.PFV.3-PL COMP  
               qal-t-i-l-hom    [li    ġħ  
               say.PFV-3SGF-EPENT.VWL-DAT-3PL COMP have-3PL.GEN  
               and-hom    j-i-xtr-u xejn. ]]]  
               3-FRM.VWL-buy.IPFV-PL nothing  
               ‘I didn’t hear that they said she told them they have to buy anything.’

<sup>22</sup> As an alternative to (106), bi-partitive negation and a positive proform (replacing *xejn* ‘nothing’ by *xi haġa* ‘something’ in (106)), is also grammatical, retaining the same interpretation).

## 8 LFG and Semitic languages

Camilleri & Sadler (2017a) argue that the n-word proforms like *xejn* are not in fact simply NCIS but have the broader distribution of weak NPIS, a view supported by the fact that they occur in a range of non-veridical contexts, as shown in (107), and unlike NCIS are not limited to negative or anti-veridical contexts. Equally, the *-x* of bipartite negation shares the wider distribution of an NPI, occurring in a range of contexts including conditionals, interrogatives, rhetorical interrogatives, embedded interrogatives and counterfactuals.

- (107) Maltese (Camilleri & Sadler 2017a: 154)  
 Kil-t            xejn      ċikkulata?  
 eat.PFV-2SG nothing chocolate  
 ‘Did you eat any chocolate?’

As part of the analysis they provide an approach to bi-partitite negation in Arabic dialects (primarily found in the dialects westward from the Levant to Morocco). There is both a dependency and an essential asymmetry in the distribution of *ma* and *-x*: *ma* realizes sentential negation but requires the presence of either *-x* or one or more NCI items within an appropriate domain, while *-x* itself is incompatible with the presence of (other) NCI items within that domain. Following Przepiórkowski & Patejuk (2015), Camilleri & Sadler (2017a) propose that *ma* introduces an ENEG feature. Because *ma* cannot stand alone it also introduces a constraining equation requiring a positive value of a NVM (for non-veridical marker) feature within an appropriate domain, which can be satisfied by a strictly local *-x* or by NC items in the N-series, within a certain domain.<sup>23</sup> The lexical entry for the sentential negation marker *ma* is in (108). The first line provides a value for the sentential negation feature ENEG, treating it as a feature with instantiated values, with the consequence that it is required to be uniquely contributed, so expressed only once. The somewhat complicated uncertainty statement requires that there either be a feature NVM = + in the local f-structure (which will be introduced by *-x*, see example (102) and the entry for *-x* in (109)) or that some dependent within the domain specified by the functional uncertainty path be specified as NVM = + (e.g. examples (104a), (106), where NVM = + is associated with an n-word dependent, see the entry for *xejn* in (110)). This path rules out *ma* satisfying its requirement for a NVM = + dependent in a subordinate negative domain, ruling out (111). The non-veridicality affix *-x* defines NVM = + and is incompatible with NVM = + on any local dependent or any more deeply embedded

<sup>23</sup>Because both *-x* and the N-series proforms occur in the wider set of non-veridical contexts they cannot simply be associated with an inside-out statement limiting them to contexts containing ENEG = +.

Louisa Sadler

dependent which is not itself inside an f-structure marked as ENG = +, thus ruling out (112). The entry for an N-series word simply defines the NVM feature in the local f-structure, as in (110).

- (108) *ma*                      ENEG     $(\uparrow \text{ENEG}) = +$   
 $\{ (\uparrow \{ \text{XCOMP} | \text{COMP} | \text{ADJ} \}^* \text{GF}^+ \text{NVM}) \mid (\uparrow \text{NVM}) \} =_{\text{C}} +$   
 $\neg(\rightarrow \text{ENEG})$   
 (Camilleri & Sadler 2017a: 159)

- (109) *-x*                       $(\uparrow \text{NVM}) = +$   
 $\neg(\uparrow \{ \text{XCOMP} | \text{COMP} | \text{ADJ} \}^* \text{GF}^+ \text{NVM}) = +$   
 $\neg(\rightarrow \text{ENEG})$   
 (Camilleri & Sadler 2017a: 159)

- (110) *xejn*                      N     $(\uparrow \text{NVM}) = +$                       Camilleri & Sadler (2017a: 159)

- (111) Maltese (Camilleri & Sadler 2017a: 159)  
 \*Ma semma                      [li    *ma*    ra-x                      [li    darb-u                      lil  
 NEG say.PFV.3SGM COMP NEG see.PFV.3SGM-X COMP injure.PFV.3-PL ACC  
 ebda    raġel.]]  
 some man  
 ‘He didn’t say that he didn’t see that they injured any man.’

- (112) Maltese (Camilleri & Sadler 2017a: 159)  
 \*It-tifla    ma    ra-t-x                      xejn.  
 DEF-girl NEG see.PFV-3SGF-X nothing  
 Intended: ‘The girl saw nothing.’

An example such as (106) will have the f-structure shown schematically in (113) (Camilleri & Sadler 2017a: 161).

- (113) 
$$\left[ \begin{array}{c} \text{ENEG} \quad + \\ \text{PRED} \quad \text{'HEAR<SUBJ,COMP>'} \\ \text{COMP} \quad \left[ \begin{array}{c} \dots \\ \left[ \begin{array}{c} \text{COMP} \quad \left[ \begin{array}{c} \text{PRED} \quad \text{'BUY<SUBJ,OBJ>'} \\ \text{OBJ} \quad \left[ \begin{array}{c} \text{PRED} \quad \text{'NOTHING'} \\ \text{NVM} \quad + \end{array} \right] \end{array} \right] \end{array} \right] \end{array} \right] \end{array} \right]$$

Alruwaili & Sadler (2018) look at negation, n-words and the combination of negation and coordination in a construction similar to the English *neither...nor* construction in the vernacular Arabic of Turaif in the Northern region of Saudi

Arabia. Turaif Arabic does not use the bipartite negation illustrated above for Maltese. Also unlike Maltese, the n-words which can occur as fragment answers, including the negative proform *māhad* ‘no one’ and the scalar focus particle *wala* ‘not even one’ can occur (*preverbally*) without the negation marker, giving rise to a negative interpretation, as shown in (114a). Hence a preverbal n-word in combination with the sentential negation marker *mā* results in a double negation reading, as in (115). Alruwaili & Sadler (2018) treat these negative arguments as contributing CNEG adopting the distinction between ENEG and CNEG introduced by Przepiórkowski & Patejuk (2015), and proposing the f-structure in (116) for (115).<sup>24</sup>

(114) Turaif Arabic (Alruwaili & Sadler 2018: 30)

- a. *māhad* ḡa                      l-yōm  
no.one come.PFV.3SGM DEF-today.SGM  
‘No one came today.’
- b. *mā* ḡa                      ʔaḥad l-yōm  
NEG come.PFV.3SGM one DEF-today  
‘No one came today.’

(115) Turaif Arabic (Alruwaili & Sadler 2018: 30)

- wala* ṭālib                      mā ḡ-a                      l-yōm  
NEG.SFP student.SGM NEG come.PFV-3SGM DEF-today  
‘Every student came today.’  
(= Not even a single student didn’t come today.)

(116)

[	PRED ‘COME<SUBJ>’	]
	ENEG +	
	SUBJ	
	[	
	ADJ {[PRED ‘TODAY’]}	

PRED ‘STUDENT’
CNEG +
NUM SG
SFOC +

(Alruwaili & Sadler 2018: 31)

The main focus of this paper is on the bipartite negative coordination marker *lā*... *wala* illustrated in (117b) (and found across many dialects of Arabic).

(117) Turaif Arabic (Alruwaili & Sadler 2018: 32–33)

---

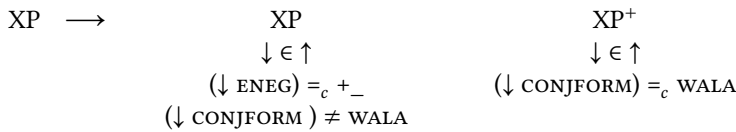
<sup>24</sup>The feature SFOC is associated with the scalar focus determiner *wala*.

Louisa Sadler

- a. mansōr mā gaʕad min n-nōm, w ʕali mā  
 Mansour NEG wake.PFV.3SGM from DEF-sleep, CONJ Ali NEG  
 ġa min d-dawām  
 come.PFV.3SGM from DEF-work  
 ‘Mansour did not wake up and Ali didn’t come (back) from work.’
- b. lā mansōr gaʕad min n-nōm, wala ʕali  
 NEG Mansour wake.PFV.3SGM from DEF-sleep, NEG.CONJ Ali  
 ġa min d-dawām  
 come.PFV.3SGM from DEF-work  
 ‘Mansour did not wake up and nor did Ali come (back) from work.’

Alruwaili & Sadler (2018) analyse both the negative conjunction *wala* (which rather transparently combines the conjunction *wa* and a negative formative) and the negative marker *lā* as elements which adjoin to (and mark) a conjunct, postulating special coordination schema for *neither.....nor* coordination – the rules in (118) and (119) (Alruwaili & Sadler 2018: 38) illustrate for sentential coordination.

(118) *Negative Coordination Schema*



- (119)  $\text{XP} \longrightarrow \begin{array}{c} \text{Neg} \\ \uparrow = \downarrow \\ (\in \uparrow) \end{array} \quad \begin{array}{c} \text{XP} \\ \uparrow = \downarrow \end{array}$

- (120) *wala* Neg  $(\uparrow \text{CONJFORM}) = \text{WALA}$  (Alruwaili & Sadler 2018: 38)  
 $(\uparrow \text{ENEG}) = +_-$   
 $((\in \uparrow) \text{CONJTYPE}) = \text{AND}$

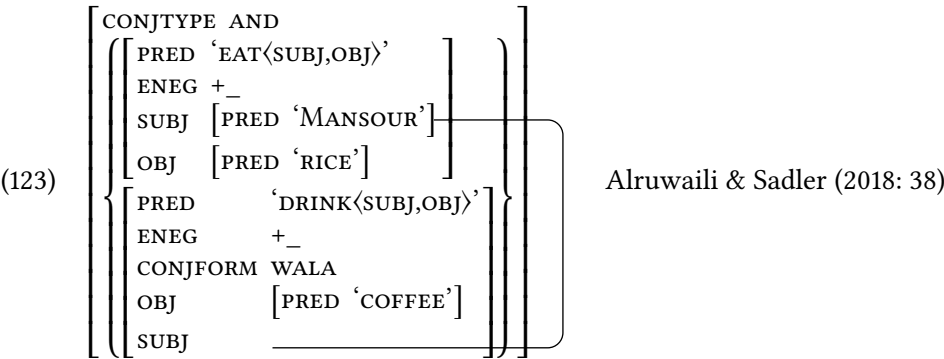
- (121) *lā* Neg  $(\uparrow \text{CONJFORM}) = \bar{\text{LĀ}}$  Alruwaili & Sadler (2018: 39)  
 $(\uparrow \text{ENEG}) = +_-$   
 $((\in \uparrow) \text{CONJTYPE}) = \text{AND}$

The f-structure for (122) on this analysis is shown in (123).

- (122) Turaif Arabic (Alruwaili & Sadler 2018: 32)



mansōr    mā akal            l-ruz    wala    šarab  
Mansour.M NEG eat.PFV.3SGM DEF-rice NEG.CONJ drink.PFV.3SGM  
l-gahwa  
DEF-coffee  
‘Mansour neither ate the rice nor drank the coffee.’



The *neither...nor* construction may also be used to coordinate arguments, where it shows the weak NCI behaviour noted above for negative elements such as *maḥad* ‘no one’ and determiner *wala*. That is, occurring preverbally, it expresses negation (and hence can give rise to double negation readings) while postverbally, it behaves like a NCI.

- (124) Turaif Arabic (Alruwaili & Sadler 2018: 34,34,40)
- a. lā    ʔaḥmad wala        mhammad    ġ-aw  
     NEG Ahmad NEG.CONJ Mohamad come.PFV-3PLM  
     ‘Neither Ahmad nor Mohammad came.’
  - b. lā    ʔaḥmad wala        mhammad    mā    ġ-aw  
     NEG Ahmad NEG.CONJ Mohammad NEG come.PFV-3PLM  
     ‘Both Ahmad and Mohammad came.’
  - c. mā    ġ-aw                lā    ʔaḥmad wala        ʕali  
     NEG come.PFV-3PLM NEG Ahmad.M NEG.CONJ Ali.M  
     ‘Neither Ahmad nor Ali came.’

In previous work, Przepiórkowski & Patejuk (2015) associate the Polish strict NCI *nikt* ‘nobody’ with an inside-out constraint requiring ENEG =+ to be defined in the appropriate containing f-structure. Building on this approach, Alruwaili & Sadler (2018) formulate a complex lexical constraint to capture the dependency between the CNEG/NCI alternation and the existence and linear position of a ENEG marker.

Louisa Sadler

## 8 Unbounded Dependency Constructions

Hebrew and Arabic both make extensive use of resumptive strategies as well as gap strategies in unbounded dependency constructions, and formalisation of the resumptive strategy for Hebrew is a major concern of Asudeh (2012), the most important reference for this section (see also Asudeh 2011). Falk (2002) also discusses the resumptive strategy for Hebrew UDCs. Camilleri & Sadler (2011) looks at restrictive relative clauses and resumption in Maltese (see also Camilleri & Sadler 2012a), building on Asudeh's approach to resumption. Further work on Maltese is descriptively oriented (Camilleri & Sadler 2016; Sadler & Camilleri 2017).

Hebrew resumptives occur in all NP positions except that of the highest subject. (125) illustrates an optional OBJ resumptive and (126) illustrates a resumptive within a complex NP island (note that there is no *wh*-item in these Hebrew relative clauses).

(125) Hebrew (Borer 1984: 220)

raʔiti ʔet ha-yeled she/ʔasher rina ʔohevet ʔoto  
saw.1SG ACC DEF-boy COMP Rina love.3SGF him  
'I saw the boy that Rina loves.'

(126) Hebrew (Borer 1984: 221)

raʔiti ʔet ha-yeled she-/asher dalya makira ʔet ha-ʔisha  
saw-I ACC DEF-boy COMP Dalya knows ACC DEF-woman  
she-ʔohevet ʔoto  
COMP-loves him  
'I saw the boy that Dalya knows the woman who loves him.'

It is well established in the literature beyond LFG that the resumptives of Hebrew have the interpretational properties of pronouns rather than those of gap. The diagnostics distinguishing those which are interpretationally identical to gaps from those which behave semantically as pronouns include differences in behaviour with respect to island phenomena, weak crossover, across-the-board extraction, parasitic gaps and reconstruction (McCloskey 2017: 106). In line with this work, Asudeh (2011, 2012) distinguishes two types of true resumptives, which he refers to as *syntactically active resumptives* (SARS) and *syntactically inactive resumptive* (SIRS). Both types of resumptive receive the same treatment in the syntax-semantics interface, that is, they are removed by a manager resource. SARS do not display gap-like properties in the syntax and are simply anaphorically bound pronouns in the syntax: the RPs of Hebrew are of this type, as shown

in (128). On the other hand, (SIRS) are syntactically gap-like (i.e. they are functionally controlled): the *RP* is treated as the bottom of a filler-gap dependency by restricting out the pronominal *PRED*, so that syntactically, the *RP* is equivalent to a gap (this analysis is given for Swedish in Asudeh 2012).

On the view that Asudeh develops, Hebrew resumptives are pronouns at *f*-structure, and are licensed in the complementiser system of Hebrew.<sup>25</sup>

That is, members of the class of *C* elements are lexically associated with the (optional) information shown in (127).

$$\begin{aligned}
 (127) \quad C \quad & \% RP = (\uparrow GF^+) & (\text{Asudeh 2012: 221}) \\
 & (\uparrow UDF)_\sigma = (\% RP_\sigma \text{ ANTECEDENT}) \\
 & @MR(\% RP) \\
 & @RELABEL(\% RP)
 \end{aligned}$$

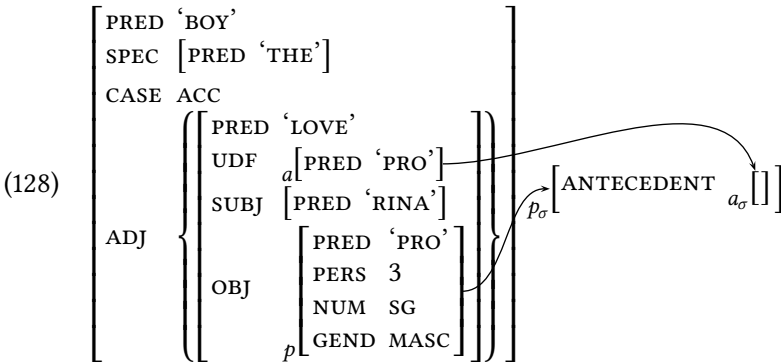
Abstracting away from many technical details, (127) states an equality between the semantics of a discourse function ( $\uparrow UDF$ ) in the *f*-structure which contains the complementiser and the value of the *ANTECEDENT* attribute of some grammatical function within the structure (identified by means of the local name *%RP*). The template call in the third line introduces the semantic resource which removes the surplus pronominal resource in the course of semantic composition, using the Resource Management Theory of Resumption developed in Asudeh (2012). The example in (125) with the resumptive has the *f*-structure in (128) (Asudeh 2012: 227).<sup>26</sup> The (standard) *CP* rule is shown in (129) (Asudeh 2012: 224) where  $\epsilon$  is not an empty node in the *c*-structure but the absence of a node associated with the collection of constraints specified.

<sup>25</sup> An alternative view of the resumptive pronouns is taken in Falk (2002), namely that pronouns may lack a *PRED* value just in case they are functionally identified with a discourse function: functional identification is introduced lexically (by the pronoun itself) and mediated by reference to a *p* projection containing the referential elements in the discourse as shown in (i).

(i)  $f \in p^{-1}(\uparrow_p) \wedge (DF f) \Rightarrow \uparrow = f$  (Falk 2002: 163)

<sup>26</sup> Asudeh does not represent the subcategorised arguments within the *PRED* value, which is a simple, argument-less semantic form.

Louisa Sadler



(129)  $CP \rightarrow \left\{ \begin{array}{c} XP \mid \epsilon \\ (\uparrow UDF) = \downarrow \quad (\uparrow UDF \text{ PRED}) = \text{'PRO'} \\ (\text{ADJUNCT} \in \uparrow) \\ REL_{\sigma} \end{array} \right\} C' \quad \uparrow = \downarrow$

Asudeh (2012) provides detailed coverage of many aspects of the syntax of Hebrew UDCs. For example (130) contains a fronted resumptive and no complementiser. The former is treated as an adjunction to C and the latter by means of a lexical entry for a null complementiser. *?ašer* is a complementiser which can only appear in relative clauses, a restriction which is captured by an inside-out constraint in the lexical entry (132)

- (130) Hebrew (Borer 1984: 220)  
 raʔiti ʔet ha-yeled ʔoto rina ʔohevet  
 saw.1SG ACC DEF-boy him Rina love.3SGF  
 'I saw the boy that Rina loves.'

(131)  $C \rightarrow \begin{array}{cc} C & \hat{D} \\ \uparrow = \downarrow & (\uparrow GF) = \downarrow \end{array} \quad (\text{Asudeh 2012: 223})$

(132) *?ašer*  $C \quad (\text{ADJUNCT} \in \uparrow) \quad (\text{Asudeh 2012: 223})$

Camilleri & Sadler (2011) provide an analysis of Maltese restrictive relative clauses. In Maltese a resumptive is not permitted in the highest subject function or, in relative clauses with definite or quantified heads, the highest object position. They suggest the underlying distribution of resumptive and gap is essentially free but subject to some additional restrictions (for example, only a resumptive is possible as the argument of a preposition).

- (133) Maltese (Camilleri & Sadler 2011: 113)

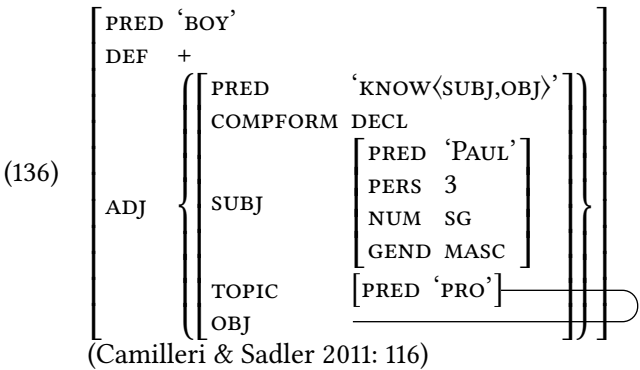
Ir-raġel li bġhatt-(lu) l-ittra  
DEF-man COMP send.PFV.1SG.(-DAT.3SGM) DEF-letter  
weġib-ni  
respond.PFV.3SGM-1SG.ACC  
'The man that I sent (him) the letter responded.'

As well as complementiser-introduced relatives such as (133), Maltese also has *wh*-relatives, which involve a gap rather than a resumptive pronoun, although these are subject to quite severe restrictions. (134) is an example.

- (134) Maltese (Camilleri & Sadler 2011: 114)  
It-tifel 'l min n(a)-hseb j-għallem-\*u  
DEF-boy ACC.who 1SG-think.IPFV 3-teach.IPFV.3SGM-3SG.ACC  
'the boy who I think he teaches'

Building on standard assumptions, Camilleri & Sadler (2011) provide a syntactic analysis of both complementiser and *wh*-relatives. The example in (135) with either a complementiser or a *wh*-item is associated with the f-structure in (136) (assuming the PRED value of 'l min is 'PRO').

- (135) Maltese (Camilleri & Sadler 2011: 116)  
Rajt lit-tifel li /'l min j-af Pawlu  
see.PFV.1SG ACC.DEF-boy COMP /who 3SGM-know.IPFV Paul  
'I saw the boy that Paul knows.'



Camilleri & Sadler (2011) show that Maltese also has true resumptives (as opposed to intrusive pronouns), and that the available tests indicate that (in the terminology of Asudeh 2012) they are SARS and hence anaphorically bound pronouns in the syntax. For example, they can be used felicitously in circumstances

Louisa Sadler

which would induce weak crossover violations. In (137) the dependency between the antecedent (*ir-raġel*) (or the TOPIC) and the RP ‘crosses over’ the possessive in *martu* (‘his wife’), but the sentence is completely grammatical, while the corresponding sentence with a gap would be ungrammatical, despite the fact that RPs are normally excluded in *wh*-relatives in Maltese. Note that the *POSS* function is not accessible to relativisation by the *wh*-strategy and so it is clear that (137) involves relativisation on the OBJ, and therefore constitutes a case of crossover. (138) provides a similar example using the less restricted complementiser strategy for relativisation.

- (137) Maltese (Camilleri & Sadler 2011: 19)

Ir-raġel 'l min n-af li t-elq-it-u  
 DEF-man ACC.who 1SG-know.IPFV COMP 3SGF-leave.PFV-3SGM.ACC  
 l-mara/mart-\*(u)  
 DEF-woman/woman-3SGM.ACC  
 ‘the man who I know that his wife left him’

- (138) Maltese (Camilleri & Sadler 2011: 19)

Ir-raġel li n-af li ħallie-t-u  
 DEF-man COMP 1.SG-know.IPFV COMP leave.PFV-3SGF-3SGM.ACC  
 mart-\*(u) baqa' ma hariġ-x mid-dar  
 wife-3SGM.ACC stay.PFV.3SGM NEG go out.3SGM-NEG from.DEF-house  
 ‘The man who I know that his wife left him, has not left the house since.’

(139) illustrates the Complex Noun Phrase Constraint, with a (second) relative dependency into a CNP created by relativisation: although the relativised position is one which is normally accessible to the gap strategy, the resumptive is obligatory here as a gap would cause a syntactic constraint violation.<sup>27</sup>

- (139) Maltese (Camilleri & Sadler 2011: 120)

Raj-t ir-raġel li n-af mara li  
 see.PFV-1SG DEF-man COMP 1SG-know.IPFV woman COMP  
 t-af-u u għid-t-l-u  
 3SGF-know.IPFV-3SGM.ACC and tell.PFV-1SG-DAT-3SGM

<sup>27</sup>The distribution of resumptives in Maltese does raise some potentially puzzling issues. Camilleri & Sadler (2011) show that there may be evidence from the distribution of gaps and RPs in across-the-board constructions that Maltese also has *syntactically inactive resumptives* (SIRS) (functionally controlled RPs or ‘audible’ gaps) since gaps and resumptives occur together in ATB constructions, but that simply assuming that ATB constructions in Maltese (and in Arabic more widely) involve SIRS rather than SARS is also problematic.

8 *LFG and Semitic languages*

j-selli-l-i

ghali-ha

3SGM-send regards.IPFV-DAT-1SG for-3SGF.ACC

‘I saw the man who I know a woman that knows him, and told him to send her my regards.’

## 9 Other Work

Alotaibi (2014) looks at conditional sentences in Hijazi Arabic and provides an LFG analysis of the syntax of these constructions. Camilleri et al. (2014a) discusses the dative alternation in Hijazi Arabic, ECA and Maltese and develops an account of the mapping to GFS using the mapping approach of Kibort (2008). Camilleri & Sadler (2012b) looks at non-selected datives in Maltese. Alzaidi (2010) on gapping constructions in Hijazi (Taif) Arabic. Sadler (2019) provides an analysis of mixed agreement in adjectival relatives in MSA. Clausal possession in Hebrew is discussed in Falk (2004). For an early discussion of agreement in MSA see Fassi Fehri (1988). Camilleri & Sadler (2017b) discusses the grammaticalisation of a progressive construction in the Arabic vernaculars from a posture verb ACT.PTCP and also provides a synchronic account of the progressive construction. Camilleri & Sadler (2018) concerns the grammaticalisation of both the universal perfect (see also Camilleri 2016) and the progressive in Arabic.

## Acknowledgements

I am grateful to the reviewers for their very helpful comments on earlier versions of this chapter.

## References

- Alhailawani, Mohammad. 2018. *Nominal structure and ellipsis in Jordanian Arabic*. London: Queen Mary University of London. (Doctoral dissertation).
- Alotaibi, Yasir. 2014. *Conditional sentences in Modern Standard Arabic and the Taif dialect*. Colchester, UK: University of Essex. (Doctoral dissertation).
- Alotaibi, Yasir, Muhammad Alzaidi, Maris Camilleri, Shaimaa ElSadek & Louisa Sadler. 2013. Psychological predicates and verbal complementation in Arabic. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '13 conference*, 6–26. Stanford, CA: CSLI Publications.
- Alruwaili, Shatha. 2019. *Negation in Turaif Arabic: Not the last word*. Colchester, UK: University of Essex. (Doctoral dissertation).

Louisa Sadler

- Alruwaili, Shatha & Louisa Sadler. 2018. Negative coordination in (Turaif) Arabic. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '18 conference*, 25–45. Stanford, CA: CSLI Publications.
- Alsharif, Ahmad. 2014. *Negation in Arabic*. Colchester, UK: University of Essex. (Doctoral dissertation).
- Alsharif, Ahmad & Louisa Sadler. 2009. Negation in Modern Standard Arabic: An LFG approach. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 5–25. Stanford, CA: CSLI Publications.
- Alzaidi, Muhammad. 2010. *Gapping and right node raising: An LFG approach*. Colchester, UK: University of Essex. (MA thesis).
- Asudeh, Ash. 2011. Towards a unified theory of resumption. In Alain Rouveret (ed.), *Resumptive pronouns at the interfaces*. Amsterdam: John Benjamins. DOI: 10.1075/lfab.5.03asu.
- Asudeh, Ash. 2012. *The logic of pronominal resumption* (Oxford Studies in Theoretical Linguistics). Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199206421.001.0001.
- Asudeh, Ash & Ida Toivonen. 2012. Copy raising and perception. *Natural Language & Linguistic Theory* 30/2. 321–380. DOI: 10.1007/s11049-012-9168-2.
- Attia, Mohammed. 2008. A unified analysis of copula constructions in LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 89–108. Stanford, CA: CSLI Publications.
- Benmamoun, Elabbas. 2000. *The feature structure of functional categories: A comparative study of Arabic dialects*. Oxford: Oxford University Press.
- Borer, Hagit. 1984. Restrictive relatives in Modern Hebrew. *Natural Language & Linguistic Theory* 2. 219–260. DOI: 10.1007/bf00133282.
- Börjars, Kersti, Khawla Ghadgoud & John Payne. 2016. Aspectual object marking in Libyan Arabic. In Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King & Stefan Müller (eds.), *Proceedings of the joint 2016 conference on Head-Driven Phrase Structure Grammar and Lexical Functional Grammar*, 125–139. Stanford, CA: CSLI Publications.
- Börjars, Kersti, Safiah Madkhali & John Payne. 2015. Masdars and mixed category constructions. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '15 conference*, 289–303. Stanford, CA: CSLI Publications.
- Bresnan, Joan, Ash Asudeh, Ida Toivonen & Stephen Wechsler. 2016. *Lexical-Functional Syntax*. 2nd edn. (Blackwell Textbooks in Linguistics 16). Malden, MA: Wiley-Blackwell.
- Camilleri, Maris. 2016. *Temporal and aspectual auxiliaries in Maltese*. Colchester, UK: University of Essex. (Doctoral dissertation).



- Camilleri, Maris, Shaimaa ElSadek & Louisa Sadler. 2014a. A cross dialectal view of the Arabic dative alternation. *Acta Linguistica Hungarica* 61(1). 3–44. DOI: 10.1556/aling.61.2014.1.1.
- Camilleri, Maris, Shaimaa ElSadek & Louisa Sadler. 2014b. Perceptual reports in (dialects of) Arabic. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 179–199. Stanford, CA: CSLI Publications.
- Camilleri, Maris & Louisa Sadler. 2011. Restrictive relative clauses in Maltese and resumption. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '11 conference*, 110–130. Stanford, CA: CSLI Publications.
- Camilleri, Maris & Louisa Sadler. 2012a. *An LFG approach to non-restrictive relative clauses in Maltese*. Research Reports in Linguistics 6. Colchester, UK: University of Essex.
- Camilleri, Maris & Louisa Sadler. 2012b. On the analysis of non-selected datives in Maltese. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 118–138. Stanford, CA: CSLI Publications.
- Camilleri, Maris & Louisa Sadler. 2016. Relativisation in Maltese. *Transactions of the Philological Society* 114(1). 117–145. DOI: 10.1111/1467-968X.12070.
- Camilleri, Maris & Louisa Sadler. 2017a. Negative sensitive indefinites in Maltese. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '17 conference*, 146–166. Stanford, CA: CSLI Publications.
- Camilleri, Maris & Louisa Sadler. 2017b. Posture verbs and aspect: A view from Vernacular Arabic. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '17 conference*, 167–187. Stanford, CA: CSLI Publications.
- Camilleri, Maris & Louisa Sadler. 2018. Schematising (morpho)syntactic change in LFG: insights from grammaticalisation in Arabic. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '18 conference*, 129–149. Stanford, CA: CSLI Publications.
- ElSadek, Shaimaa. 2016. *Verbal complementation in Egyptian Colloquial Arabic: An LFG account*. Colchester, UK: University of Essex. (Doctoral dissertation).
- ElSadek, Shaimaa & Louisa Sadler. 2015. Egyptian Arabic perceptual reports. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '15 conference*, 84–102. Stanford, CA: CSLI Publications.
- Falk, Yehuda N. 2001. Constituent structure and grammatical functions in the Hebrew action nominal. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '01 conference*. Stanford, CA: CSLI Publications.
- Falk, Yehuda N. 2002. Resumptive pronouns in LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '02 conference*, 154–173. Stanford, CA: CSLI Publications.

Louisa Sadler

- Falk, Yehuda N. 2004. The Hebrew present tense copula as a mixed category. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*, 226–246. Stanford, CA: CSLI Publications.
- Falk, Yehuda N. 2006. *Subjects and universal grammar: An explanatory theory*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/cbo9780511486265.
- Falk, Yehuda N. 2007. Constituent structure and GFS in the Hebrew nominal phrase. 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*, 103–126. Stanford, CA: CSLI Publications.
- Fassi Fehri, Abdelkader. 1988. Agreement in Arabic, binding and coherence. In Michael Barlow & Charles A. Ferguson (eds.), *Agreement in natural language*, 107–158. Stanford, CA: CSLI Publications.
- Fassi Fehri, Abdelkader. 1993. *Issues in the structure of Arabic clauses and words*. Dordrecht: Kluwer Academic Publishers. DOI: 10.1007/978-94-017-1986-5.
- Hallman, Peter. 2018. Double-object constructions in Syrian Arabic. *Syntax* 21(3). 238–274. DOI: 10.1111/synt.12157.
- Kibort, Anna. 2008. On the syntax of ditransitive constructions. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 312–332. Stanford, CA: CSLI Publications.
- Kifle, Nazareth Amlesom. 2007. Differential object marking and topicality in Tigrinya. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '07 conference*, 5–25. Stanford, CA: CSLI Publications.
- Kifle, Nazareth Amlesom. 2011. *Tigrinya applicatives in Lexical-Functional Grammar*. University of Bergen.
- Lowe, John J. 2020. Mixed projections and syntactic categories. *Journal of Linguistics* 56(2). 315–357. DOI: 10.1017/S0022226719000100.
- McCloskey, James. 2017. Resumption. In Martin Everaert & Henk van Riemsdijk (eds.), *The Wiley Blackwell companion to syntax*, 2nd edn. Hoboken: John Wiley & Sons. DOI: 10.1002/9781118358733.wbsyncom105.
- Ouhalla, Jamal. 1993. Negation, focus and tense: The Arabic *maa* and *laa*. *Rivista di Linguistica* 5(2). 275–300.
- Ouwayda, Sarah. 2012. On construct state nominals: Evidence for a predicate approach. In Ghil'ad Zuckermann (ed.), *Burning issues in Afro-Asiatic linguistics*, 75–90. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Przepiórkowski, Adam & Agnieszka Patejuk. 2015. Two representations of negation in LFG: Evidence from Polish. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '15 conference*, 322–336. Stanford, CA: CSLI Publications.

## 8 LFG and Semitic languages

- Sadler, Louisa. 2000. Noun phrase structure in Welsh. In Miriam Butt & Tracy Holloway King (eds.), *Argument realization*, 73–109. Stanford, CA: CSLI Publications.
- Sadler, Louisa. 2019. Multiple controllers in nominal modification. *Argumentum* 15. 617–638.
- Sadler, Louisa & Maris Camilleri. 2017. Free relatives in Maltese. *Brill's Journal of Afroasiatic Languages and Linguistics* 10. 115–159. DOI: 10.1163/18766633-00901001.
- Sichel, Ivy. 1997. Two pronominal copulas and the syntax of Hebrew nonverbal sentences. In Ralph C. Blight & Michelle Moosally (eds.), *Proceedings of the 1997 Texas Linguistics Society Conference*, 295–306.
- Toivonen, Ida. 2003. *Non-projecting words: A case study of Swedish particles*. Dordrecht: Kluwer Academic Publishers.
- Wintner, Shuly. 2000. Definiteness in the Hebrew noun phrase. *Journal of Linguistics* 36. 319–363. DOI: 10.1017/s0022226700008173.



## **Part V**

# **Comparing LFG with other linguistic theories**



# Chapter 9

## LFG and HPSG

Adam Przepiórkowski

University of Warsaw; Polish Academy of Sciences; University of Oxford

This chapter presents and compares Lexical Functional Grammar and Head-driven Phrase Structure Grammar. It concentrates on their fundamental properties rather than on analyses of particular phenomena. After discussing representations assumed in each theory and the kinds of grammars that lead to such representations, the chapter devotes some attention to models explicitly or implicitly assumed in HPSG and LFG: it identifies some problems and suggests possible solutions.

### 1 Introduction

The aim of this chapter is to juxtapose two highly formalised grammatical theories: Lexical Functional Grammar (LFG; Kaplan & Bresnan 1982; Bresnan et al. 2016; Dalrymple et al. 2019) and Head-driven Phrase Structure Grammar (HPSG; Pollard & Sag 1987, 1994; Müller et al. 2021).<sup>1</sup> LFG was conceived in the late 1970s, HPSG – in the mid-1980s, so both theories have been around for decades. Within both theories, diverse phenomena have been analysed and then re-analysed, and many will undoubtedly receive new analyses in the future. For this reason, rather than compare particular analyses of some phenomena, this chapter focuses on more fundamental issues: on the general representational architecture of the two theories (in Section 2), on the kinds of grammars that lead to these representations (in Section 3), and on models assumed in both theories (in Section 4). Wechsler & Asudeh (2021) offers a comparison of the treatment of various phenomena in the two theories and, hence, complements the current chapter.

---

<sup>1</sup>This chapter does not presuppose substantial prior exposure to either LFG or HPSG.



2 Representations

Outside of their respective communities, both theories are best known as theories of syntax, although already at their conception both were envisaged as theories of multiple linguistic levels, including semantics. Current versions of both theories have well-developed approaches to semantics, as well as proposals for the representation of other linguistic levels: morphological and information-structural in the case of both theories, phonological in the case of HPSG, and prosodic in the case of LFG.

However, the two theories adopt rather different approaches to the representation of the various linguistic levels.

2.1 LFG

Let us have a look at possible representations of the simple sentence (1) in the two theories, starting with the LFG representation in Figure 1.

(1) She loves you.

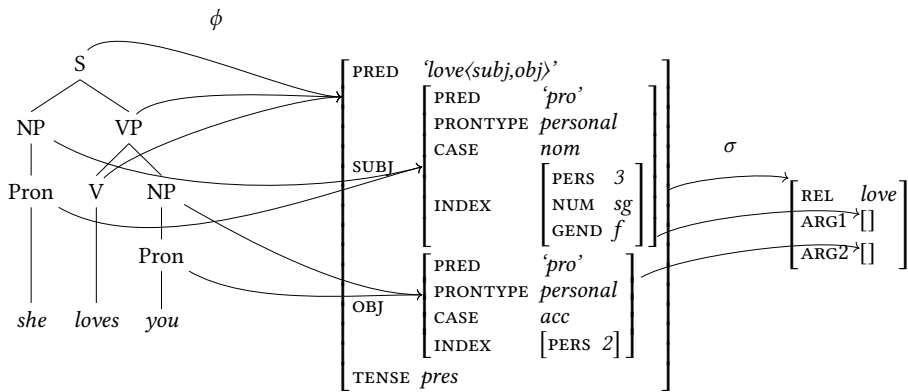


Figure 1: LFG representation of (1)

A prominent feature of LFG representations are the multiple levels. Figure 1 features three such levels: constituent structure (c-structure; the tree on the left), functional structure (f-structure; the attribute–value matrix, AVM for short, in the middle), and semantic structure (s-structure; the AVM on the right). The first two of these levels are syntactic in nature and they are the core of any LFG analysis. The repertoire and exact properties of other levels, including s-structure, is



a matter of some debate. Among other prominent levels widely assumed in LFG are prosodic structure (for overviews see Dalrymple et al. 2019: Chapter 11 and Bögel 2021 [this volume]) and information structure (see Dalrymple et al. 2019: Chapter 10 and **chapters/InformationStructure**). Also argument structure used to be assumed as a separate level (see, e.g., Butt et al. 1997), but given an appropriately spelled-out approach to semantics, a separate a-structure does not seem to be needed (see, e.g., Asudeh & Giorgolo 2012 and Findlay 2016).

As shown in Figure 1, levels of representation are connected via mapping functions (rendered in the figure with arrows between levels). One such function, usually called  $\phi$ , maps c-structures to f-structures, another,  $\sigma$ , maps f-structures to s-structures. These functions are not necessarily total. In particular, it is often assumed that  $\phi$  maps to f-structures only non-terminal nodes of c-structures. For example, in Figure 1, the leftmost nodes NP and Pron, but not the terminal node *she* that they dominate, map to the f-structure representing the subject (the value of the SUBJ attribute), the rightmost nodes NP and Pron, but not the terminal *you*, map to the f-structure representing the object, etc. Similarly, the domain of  $\sigma$  consists of three f-structures (the ones containing the PRED attribute), with the exclusion of the f-structures which are the values of INDEX. These functions are also not surjective (not onto), for example the values of INDEX in the f-structure are not in the range of  $\phi$ .

Let us take a brief look at particular levels. The c-structure in Figure 1 should be self-explanatory. Unlike derivational theories (see **chapters/Minimalism**), but like Simpler Syntax (see **chapters/SimplerSyntax**) and HPSG, LFG assumes very simple constituency trees, usually without empty categories – but see Bresnan et al. (2016: Chapter 9) for exceptions – and without an abundance of functional nodes. Constituency structures are assumed to vary considerably between languages, even though their grammars are required to follow some – appropriately relaxed – version of the X'-theory.<sup>2</sup>

On the other hand, functional structures are cross-linguistically more uniform. While they contain morphosyntactic information, which is quite different for different languages, their main function is to represent grammatical functions such as subject and object, and the repertoire of grammatical functions is supposed to be universal.<sup>3</sup> F-structures also contain “semantic forms” – values of

<sup>2</sup>See, e.g., Bresnan et al. 2016: Chapter 6 and Dalrymple et al. 2019: Section 3.2. LFG versions of X'-theory are relaxed in various ways. While the standard X'-theory assumes at most binary branching, LFG does not make such an assumption. Also, standard derivational versions of X'-theory assume the presence of the head (perhaps subsequently moved to a different tree position or realised as a phonetically empty constituent to start with), while in LFG the head may be optional in a rule and completely absent from the resulting tree. In this sense, LFG versions of X'-theory may be construed as theories of descriptions rather than structures.

<sup>3</sup>See Patejuk & Przepiórkowski (2016) for a critical discussion of this assumption and Kaplan (2017) for a reply.

*Adam Przepiórkowski*

the PRED attribute – originally designed to encode in syntax the information that maps to semantic representations; as repeatedly noted in the literature, this information is largely redundant in contemporary LFG, given the existence of semantic structures.<sup>4</sup> In the case of Figure 1, the main f-structure represents a present-tense utterance with the semantic form ‘LOVE<SUBJ,OBJ>’. Both the subject and the object of this utterance have the semantic form ‘PRO’, i.e., they are pronouns, specifically, personal pronouns. Their morphosyntactic information is represented within the values of CASE and INDEX.<sup>5</sup>

Finally, s-structures contain purely semantic information. In the case at hand, it is the information that the meaning of this utterance is modelled by the relation LOVE and that there are two arguments of this relation, corresponding to the subject and the object.

## 2.2 HPSG

Let us now have a look at the HPSG representation of (1) in Figure 2. HPSG representations are formally more uniform: there is just one contiguous data structure used for the representation of information from all linguistic levels, namely, an attribute–value matrix.<sup>6</sup> In particular, there are no separate levels of representation – all constituency, morphosyntactic, and semantic information is interspersed throughout the structure.

Clearly, the cost of the greater formal uniformity is the diminished perspicuity (or, for an unaccustomed eye, downright unreadability) of representations such as that in Figure 2. For this reason, it is common among HPSG practitioners to use all kinds of abbreviations and representational devices to make representations more readable. For example, the structure of that figure may be presented as in Figure 3, where the constituency structure becomes transparent.

Taking a closer look at the AVM in Figure 2 we may first note that, unlike f-structures (or s-structures) in LFG, feature structures in HPSG are typed. For example, the structure represented by the whole AVM is of type *hd-subj-ph* (i.e., *head-subject-phrase*), and the value of the attribute HD-DTR is of type *hd-comp-ph* (i.e., *head-complement-phrase*).

<sup>4</sup>See, e.g., Dalrymple et al. (1993: 13–14) and Kuhn (2001: Sections 1.3.3, 1.4.1).

<sup>5</sup>Analyses in both LFG and HPSG often follow Wechsler & Zlatić (2003) and distinguish between INDEX agreement and CONCORD agreement; here I retain INDEX as a separate bundle of features but do not explicitly represent the CONCORD bundle, just the CASE feature within it.

<sup>6</sup>Figure 2 also contains lists, indicated with angle brackets, but this is a shorthand notation for AVMs with attributes such as FIRST and REST (or HEAD and TAIL), whose values are the first element (head) of the list and the rest (tail) of the list.

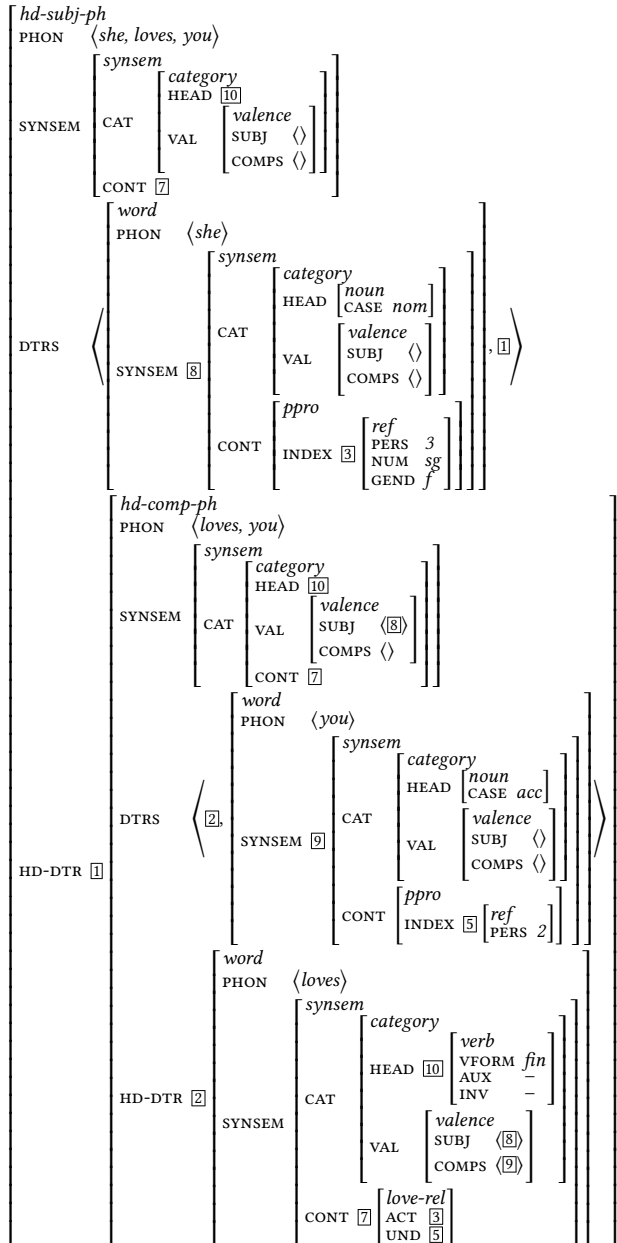


Figure 2: HPSG representation of (1)

Adam Przepiórkowski

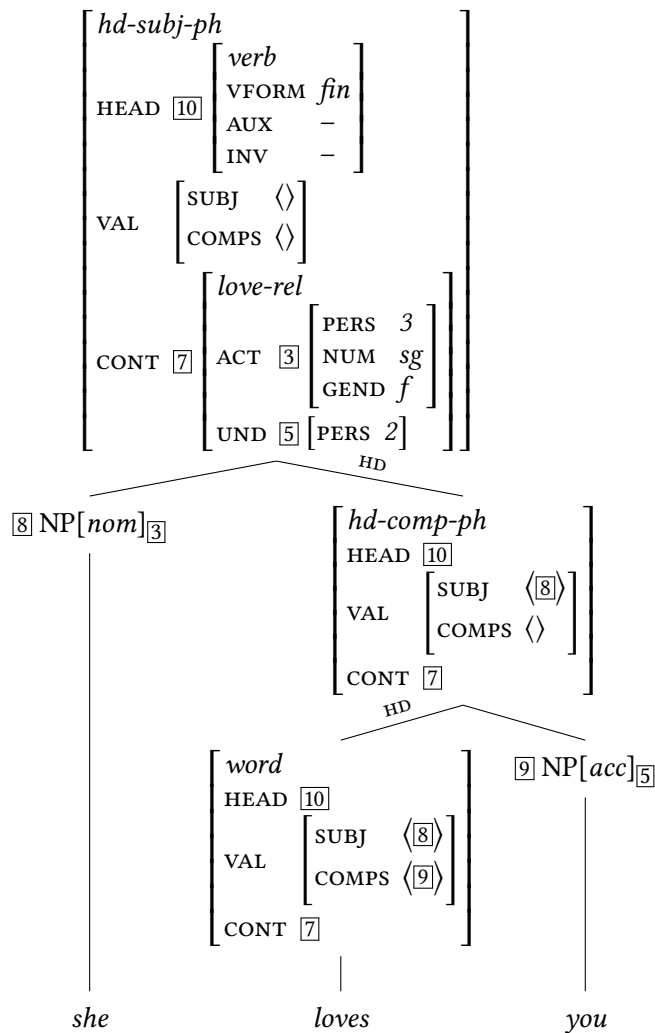


Figure 3: Shorthand HPSG representation of (1)

As discussed in more detail in Section 3.2 below, types determine what attributes may and must appear on the objects described by the AVM (not necessarily on the AVM itself, which may be a partial description of such objects; this point will be crucial below) and what their values may and must be.<sup>7</sup> Types are ordered in an inheritance hierarchy, where subtypes inherit conditions imposed by supertypes and may add more such conditions.<sup>8</sup> For example, both *hd-subj-ph* and *hd-comp-ph* are subtypes of *headed-phrase*, which is a subtype of *phrase*, which in turn – along with *word* – is a subtype of *sign*; see Figure 4.

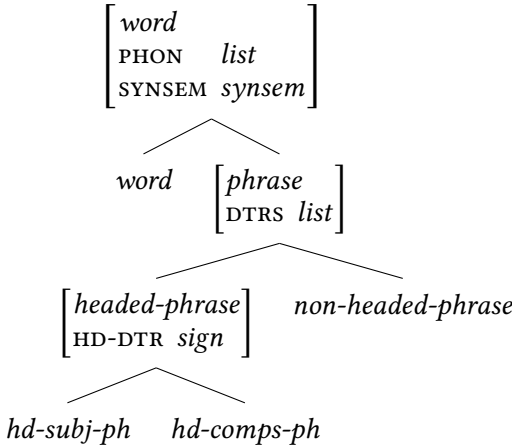


Figure 4: A small fragment of an HPSG type hierarchy

All objects of type *sign* must have two attributes: PHON and SYNSEM (I will explain their role shortly). The *word* subtype does not add any conditions, and all the three subsidiary AVMs of type *word* in Figure 2 have exactly these two attributes and no others. On the other hand, the *phrase* subtype of *sign* requires an additional attribute, namely, DTRS (i.e., DAUGHTERS), whose value is a list of immediate constituents. An important subtype of *phrase* is *headed-phrase*, where one of the immediate constituents is singled out as the syntactic head; this constituent

<sup>7</sup>In the HPSG lingo, this amounts to saying that feature structures are *totally well-typed* (Carpenter 1992: 94–95; Pollard & Sag 1994: 18).

<sup>8</sup>It is sometimes argued that LFG templates (which are, essentially, possibly parameterised macros, as in programming languages) “can play the same role in capturing linguistic generalizations as hierarchical type systems in theories like HPSG” (Dalrymple et al. 2004: 207); unfortunately, a discussion of similarities and differences between the two mechanisms – especially, the crucial ontological differences – is outside the scope of this fairly introductory chapter.

Adam Przepiórkowski

is the value of the additional HD-DTR (i.e., HEAD-DAUGHTER) attribute. Hence, any object of type *headed-phrase* must have four attributes: PHON, SYNSEM, DTRS, and HD-DTR. As *hd-subj-ph* and *hd-comp-ph* do not add any attributes, the two AVMs corresponding to the phrases *she loves you* and *loves you* have exactly these four attributes.

Let us take a closer look at the encoding of constituency structure via the attributes DTRS and HD-DTR. In the root AVM of Figure 2, the value of DTRS is a 2-element list, whose first element is a *word* structure of *she* and the second element is a *hd-comp-ph* structure of *loves you*. This second element is only marked as  $\boxed{1}$  on the DTRS list, but it is fully presented as the value of the HD-DTR attribute; boxed numbers such as  $\boxed{1}$  should be understood as bound variables signalling multiple occurrences of a structure in different places (here, in the DTRS list and in the value of HD-DTR). The structure  $\boxed{1}$ , being (a subtype of) a headed phrase, also has the attribute DTRS, whose value is a pair of structures of *loves* and *you*, and the HD-DTR attribute, which singles out the structure of *loves* as the head. This configuration of attributes DTRS and HD-DTR and their values encodes the syntactic tree of Figure 3.

The other two attributes of *phrase* structures, present also on *word* structures, are PHON and SYNSEM. In work which does not deal with phonology or phonetics the values of PHON are taken to be lists of words, as in Figure 2, but it is clear that in an exhaustive representation values of PHON must be highly structured.<sup>9</sup>

For our purposes, values of SYNSEM are more important – they represent all grammatical information other than constituency structure. Figure 2 presents slightly simplified values of SYNSEM: it omits those parts of *synsem* structures which are responsible for non-local information, i.e., for book keeping related to unbounded dependencies and relative clauses (see Borsley & Crysmann 2021, Arnold & Godard 2021, Chaves 2021, and references therein).<sup>10</sup> Local information is distributed between the attributes CAT(egory) and CONT(ent), as well as CONTEXT, not represented here either (see Pollard & Sag 1994: 332–337, as well as De Kuthy 2021 and references therein). CONT represents semantic information comparable to that distributed between LFG f-structures and s-structures. For example, the two personal pronouns (see the two CONT values of type *ppro*) contribute referential indices, referred to as  $\boxed{3}$  and  $\boxed{5}$ , and the verb contributes the *love-rel(ation)* with the index  $\boxed{3}$  of *she* as its ACT(or) and the index  $\boxed{5}$  of *you* as

<sup>9</sup>See, e.g., Bird & Klein (1994) and Höhle (1999) for two very different proposals.

<sup>10</sup>Normally, *synsem* structures contain two attributes, LOCAL and NONLOCAL. Since, NONLOCAL and its value is omitted here, also the attribute LOCAL is not mentioned in this chapter, and its values of type *local* are presented as SYNSEM values of type *synsem*.

its `UND(ergoer)`. This verbal semantics is shared along the verbal spine, so the structures of *loves*, *loves you*, and *she loves you* all have the same `CONT` value [7].

The other part of `SYNSEM` values, the *category* structure, models morphosyntactic and combinatorial properties. The former are the value of `HEAD`: *she* is a nominative (pro)noun, *you* is (here) an accusative (pro)noun, and *loves* is a finite verb (non-auxiliary, not inverted). The values of `HEAD` are shared between a mother and its head daughter – see the multiple occurrences of [10]. Finally, combinatorial properties are encoded in values of `VAL(ence)`: the verb *loves* requires a subject ([8] – the `SYNSEM` value of *she*) and a complement ([9] – the `SYNSEM` value of *you*), *loves you* has no further complement expectations but still needs a subject, while *she loves you* is a fully saturated maximal projection – the values of its valency features are empty lists. Such maximal projections are often abbreviated the way illustrated in Figure 3: `NP[nom]3` stands for (the `SYNSEM` value of) a structure with empty `SUBJ` and `COMPS`, with `HEAD` indicating a nominative noun, and with `CONT|INDEX` value [3].<sup>11</sup>

## 2.3 Comparison

### 2.3.1 Levels of representation

The two structures in Figures 1 and 2 look somewhat similar in the sense that they both use complex AVMs, but also very different in the sense that the LFG representation distinguishes multiple levels, each with its own data structure and with a functional mapping between the levels, while the HPSG representation is a monolithic AVM. How important is this difference? My claim is that it is less important than usually assumed. For example, it is possible to define a bijection (a one-to-one correspondence) between LFG representations such as that in Figure 1 and corresponding HPSG-like monolithic AVM representations such as that in Figure 5. In this representation, the c-structure is encoded with the help of attributes `LABEL`, `DTRS`, and `PHON`, the mapping  $\phi$  from the c-structure to the f-structure is achieved with the attribute `SYNSEM`, and the mapping  $\sigma$  from the f-structure to the s-structure – with the attribute `CONT`.<sup>12</sup>

Conversely, the HPSG representation of Figure 2 might be taken apart and LFG-ified as in Figure 6. The fact that non-terminal nodes in the c-structure are AVMs is not a problem in itself; in LFG it is often assumed that c-structure node labels

<sup>11</sup>While in LFG the attribute separator in paths is a space, e.g., “`SUBJ CASE`”, in HPSG the vertical bar is used, e.g., “`SYNSEM|CAT|HEAD|CASE`”.

<sup>12</sup>In fact, this representation makes conspicuous the redundancy – mentioned in Section 2.1 – of `PRED` values with respect to s-structures (i.e., here, `CONT` values).

Adam Przepiórkowski

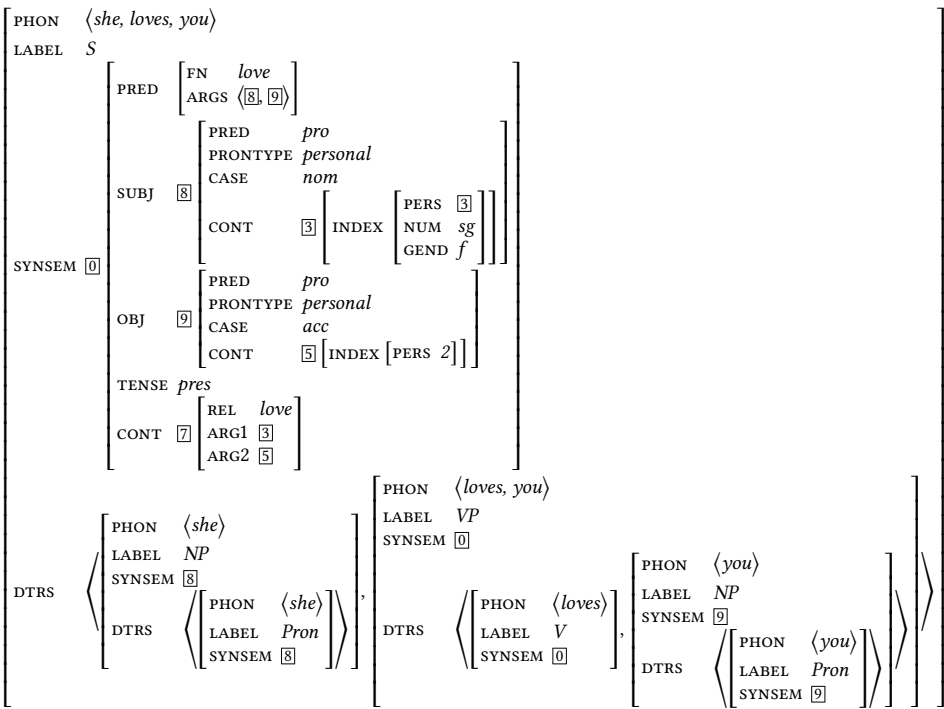


Figure 5: HPSG-like LFG representation of (1)

are really abbreviations of feature matrices (see, e.g., Kaplan 1995, Dalrymple 2017, and Lowe & Lovestrand 2020). What is somewhat unusual is that some of the attributes in these AVMs are list-valued and refer to other AVMs within the same c-structure (rather than to particular values within such AVMs as in, e.g., Lowe & Lovestrand 2020). However, this does not seem to violate any deep LFG principles.

What LFG grammars and the multi-level representations they lead to try to capture is the cognitive modularity and encapsulation of particular linguistic levels; constituency structures, functional structures, semantic structures, etc., each have their own sets of primitives and operations, and the interactions between them are only possible via the mapping functions  $\phi$ ,  $\sigma$ , etc. By contrast, no such encapsulation is attempted in HPSG, so it is easy to state constraints in this theory which may simultaneously refer to arbitrary parts of the structure of a sentence, e.g., the phonetic properties of a verb and the semantics of its subject; such a constraint would be much more cumbersome to state in LFG. On the other hand, actual LFG analyses sometimes make use of the converses of  $\phi$ ,



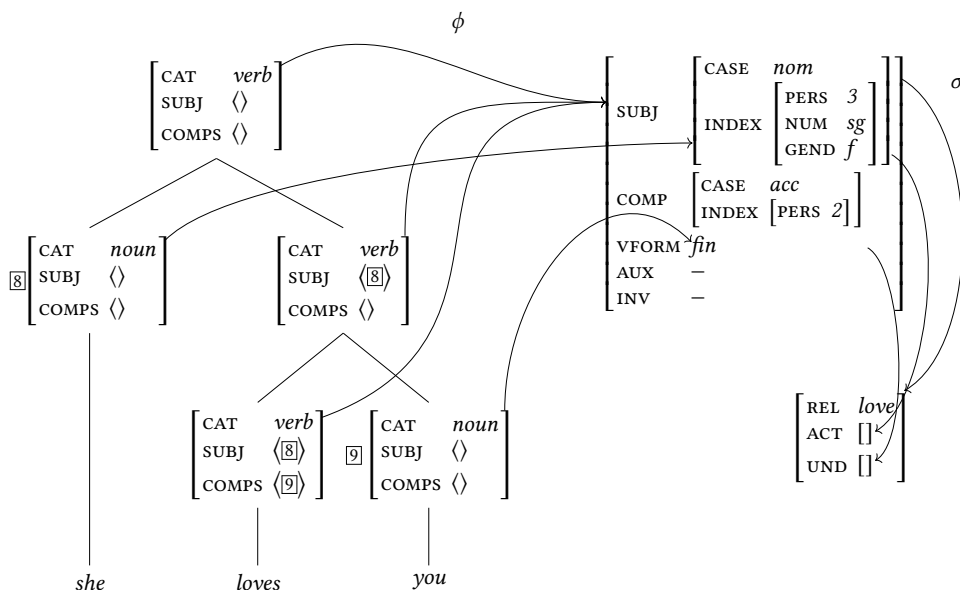


Figure 6: LFG-like HPSG representation of (1)

$\sigma$ , etc., i.e., refer to c-structures from the level of f-structures and to f-structures from the level of s-structures, so, in principle, any level may be referred to from any other level.<sup>13</sup> Hence, the difference between LFG and HPSG when it comes to encapsulation of linguistic levels is one of degree – and relative easiness of stating constraints across grammatical levels – rather than a categorical difference between the complete encapsulation and the total lack thereof.

In summary, while representations with separate linguistic levels such as those in Figures 1 and 6 are certainly more immediately readable than monolithic representations such as those in Figures 2 and 5, it is not clear that there are any fundamental differences in the kinds of linguistic analyses that LFG and HPSG presuppose.<sup>14</sup>

<sup>13</sup>However, as pointed out by Ash Asudeh (p.c.), correspondence functions in LFG are typically not injective (i.e., they are many-to-one), so their converses are proper relations rather than functions. For example, while  $\phi$  maps particular c-structure nodes to particular f-structures, the converse of  $\phi$  will map f-structures to sets of c-structure nodes, making it more difficult to refer to particular c-structure nodes from the level of f-structures. This “blurring” or “fuzziness” of converses of correspondence functions might be claimed to constitute a substantive hypothesis about encapsulation of grammatical levels.

<sup>14</sup>But see Section 3.3.3, on the expressiveness of formalisms underlying LFG and HPSG.

Adam Przepiórkowski

### 2.3.2 Grammatical functions

Perhaps a more important – and certainly linguistically more contentful – difference between HPSG and LFG regards grammatical functions. In LFG each argument bears a different grammatical function drawn from a repertoire that includes SUBJ(ect) and OBJ(ect), as in Figure 1, but also OBL(ique), COMP(lement) – a closed sentential argument, XCOMP – an open verbal argument, etc. Moreover, at least OBJ and OBL are often indexed with thematic roles, grammatical cases, or particular prepositions. For example, in the case of sentence (2), the f-structure would contain another attribute apart from SUBJ (for *you*) and OBJ (for *me*), namely, OBJ<sub>THEME</sub> (for *your money*). Similarly, in the case of (3), the grammatical function of *to you* could be OBL<sub>GOAL</sub>, etc. (see, e.g., Dalrymple et al. 2019: Section 10.3 and references therein).

- (2) You never give me your money.  
 (3) But what I've got I'll give to you.

The HPSG approach to naming arguments is radically different: normally only the SUBJ(ect) is distinguished (see Pollard & Sag 1994: Chapter 9 and references therein), often for solely tree-configurational reasons, and all the other arguments are listed within the predicate's COMP(lement)s value. In the case of a 2-argument verb such as *love* this difference is not conspicuous, but in the case of, say, *give*, the two non-subject arguments would be elements of the COMPS list, whether they are realised as a direct object and a theme object, as in (2), or as a direct object and goal oblique, as in (3). Hence, the two attributes, SUBJ and COMPS, suffice for any configuration of arguments.<sup>15</sup>

Note that this is a difference between LFG and HPSG *qua* linguistic theories, not *qua* linguistic formalisms. Either approach can be simulated in the other formalism. For example, within LFG, Alsina (1996) proposes to constrain explicitly named grammatical functions to subject and object, and Patejuk & Przepiórkowski (2016) and Przepiórkowski (2016) further justify this approach and provide an LFG formalisation inspired by HPSG analyses of extended argument structure.<sup>16</sup>

<sup>15</sup> Also, the SUBJ/COMPS dichotomy is not assumed in some versions of HPSG (including the early versions of Pollard & Sag 1987 and Pollard & Sag 1994: Chapters 1–8, as well as the Sign-Based Construction Grammar of Sag 2012, sometimes perceived as a version of HPSG) and in HPSG grammars of some languages (e.g., German; Stefan Müller, p.c.).

<sup>16</sup> Two further – more formal – arguments for HPSG-like representations of grammatical functions in essentially LFG settings may be found in Johnson (1988: Chapter 4): first, they obviate the need for the LFG principles of completeness and coherence (cf. Section 3.3.4), which are encoded via formally problematic (cf. Section 4.3.3) constraining statements, and second, they

Conversely, explicit information about grammatical functions of particular arguments could be added to HPSG representations, as in Ackerman & Webelhuth (1998) or Hellan (2019).

### 2.3.3 Word forms

The final difference between the two representations in Figures 1 and 2 that I would like to point out concerns the place of the word string in these representations. Traditionally, in LFG the sequence of word forms – the form of the utterance – is the yield of the c-structure, i.e., the sequence of leaves. So finding an LFG representation of an utterance amounts to finding a grammatical representation in which the yield of the c-structure is that utterance.

On the other hand, in HPSG the sequence of words in an utterance is the value of that utterance's `PHON` attribute. This means that finding an HPSG representation of an utterance boils down to finding a grammatical structure in which the value of `PHON` is the list of words in that utterance. Normally this amounts to the same sequence of words as that read off the leaves of the constituency tree. For example, if – in a simple binary tree – the `PHON` of the first constituent is `<come>` and the `PHON` of the second is `<together>`, then the `PHON` of the mother is `<come, together>` rather than `<together, come>` (or `<drive, my, car>`, or whatever). This correspondence is explicitly present in the representation in Figure 2 and implicitly assumed in the shorthand representation in Figure 3, but there is a well-developed linearisation theory in HPSG which allows for controlled violations to this correspondence.

I will have more to say about the exact role of the string of word forms in both linguistic theories in Section 3.3.1.

## 2.4 Summary

Let us take stock of similarities and differences between the kinds of representations assumed in LFG and HPSG. The celebrated difference between the multi-level architecture of LFG and the monolithic structures assumed in HPSG is certainly important to many practitioners of both theories and has an impact on readability (of LFG representations) and the need to apply additional conventions and abbreviations (to render HPSG representations), but in my view it is

---

lead to an analysis of Dutch infinitive constructions which, unlike the standard – at that time – LFG analysis, does not violate the offline parsability constraint (cf. Section 3.3.3). (Some problems with Johnson's (1988) own analysis of Dutch infinitive constructions are pointed out in Zaenen & Kaplan 1995.)

*Adam Przepiórkowski*

of little substantial consequence. It is trivial to devise a lossless conversion of LFG representations to HPSG-like representations, and also HPSG structures may be converted to LFG-like representations which distinguish between constituency structures, structures representing other syntactic information, and semantic structures.

However, there are at least two more substantial differences conspicuous in the representations in Figures 1 and 2. One concerns grammatical functions: one function per argument in LFG and just one distinguished argument in HPSG. The other concerns the place of the sequence of words which make up an utterance: in LFG this sequence is commonly assumed to correspond to the sequence of leaves in the c-structure, while HPSG allows for dissociation between the string of words and the constituency structure.

### 3 Grammars

What kinds of grammars lead to representations such as those in Figures 1 and 2? I will first consider LFG, then HPSG, and then I will compare the two approaches.

#### 3.1 LFG

Here is the relevant part of an LFG grammar that produces the structures in Figure 1.<sup>17</sup>

*Grammar rules:*

- $$\begin{array}{ll}
 (4) \quad S & \longrightarrow \quad \begin{array}{cc} \text{NP} & \text{VP} \\ (\uparrow \text{SUBJ}) = \downarrow & \uparrow = \downarrow \\ (\downarrow \text{CASE}) = \text{NOM} & (\downarrow \text{TENSE}) \end{array} \\
 (5) \quad \text{VP} & \longrightarrow \quad \begin{array}{cc} \text{V} & \text{NP} \\ \uparrow = \downarrow & (\uparrow \text{OBJ}) = \downarrow \\ & (\downarrow \text{CASE}) = \text{ACC} \end{array} \\
 (6) \quad \text{NP} & \longrightarrow \quad \begin{array}{c} \text{Pron} \\ \uparrow = \downarrow \end{array}
 \end{array}$$

<sup>17</sup>Only the core machinery is assumed here, mostly (apart from the  $\sigma$ -projected s-structures) present already in Kaplan & Bresnan (1982). See, e.g., Dalrymple et al. (2019: Chapter 6), for later additions such as functional uncertainty (including inside-out functional uncertainty), off-path constraints, the restriction operator, local names, templates, etc.

*Lexicon:*

- |                  |      |  |
|------------------|------|--|
| (7) <i>loves</i> | V    | $(\uparrow \text{ PRED}) = \text{'LOVE'}\langle \text{SUBJ}, \text{OBJ} \rangle$<br>$(\uparrow \text{ SUBJ INDEX PERS}) =_c 3$<br>$(\uparrow \text{ SUBJ INDEX NUM}) =_c \text{SG}$<br>$(\uparrow \text{ TENSE}) = \text{PRS}$<br>$(\uparrow_\sigma \text{ REL}) = \text{LOVE}$<br>$(\uparrow_\sigma \text{ ARG1}) = (\uparrow \text{ SUBJ})_\sigma$<br>$(\uparrow_\sigma \text{ ARG2}) = (\uparrow \text{ OBJ})_\sigma$ |
| (8) <i>she</i>   | Pron | $(\uparrow \text{ PRED}) = \text{'PRO'}$<br>$(\uparrow \text{ PRONTYPE}) = \text{PERSONAL}$<br>$(\uparrow \text{ CASE}) = \text{NOM}$<br>$(\uparrow \text{ INDEX PERS}) = 3$<br>$(\uparrow \text{ INDEX NUM}) = \text{SG}$<br>$(\uparrow \text{ INDEX GEND}) = \text{F}$   |
| (9) <i>you</i>   | Pron | $(\uparrow \text{ PRED}) = \text{'PRO'}$<br>$(\uparrow \text{ PRONTYPE}) = \text{PERSONAL}$<br>$(\uparrow \text{ INDEX PERS}) = 2$   |

LFG grammars may be viewed as Context-Free Grammars (CFGs) with annotations; the purely CFG part of the grammar in (4)–(9) is this:

- (4')  $S \longrightarrow \text{NP VP}$   
(5')  $\text{VP} \longrightarrow \text{V NP}$   
(6')  $\text{NP} \longrightarrow \text{Pron}$   
(7')  $\text{V} \longrightarrow \textit{loves}$   
(8')  $\text{Pron} \longrightarrow \textit{she}$   
(9')  $\text{Pron} \longrightarrow \textit{you}$

Within annotations,  $\uparrow$  refers to the f-structure associated (via the  $\phi$  function) with the mother node in the tree (i.e., with the preterminal node, in the case of lexical entries), and  $\downarrow$  refers to the f-structure associated with the current node. For example, the functional equation  $(\uparrow \text{ SUBJ}) = \downarrow$  under the NP in rule (4) for S says that the f-structure associated with the S node has the SUBJ attribute whose value is the f-structure associated with the NP node. The other equation under the NP,  $(\downarrow \text{ CASE}) = \text{NOM}$ , says that the f-structure for this NP has CASE with value NOM. The head equation  $\uparrow = \downarrow$  under the VP in the same rule says that S and VP are associated with the same f-structure.

Adam Przepiórkowski

These are so-called “defining equations” – they may be thought of as constructively building representations. The statement ( $\downarrow$  TENSE) under VP in rule (4) is a constraining condition requiring the presence of the TENSE attribute within the f-structure associated with the VP node. This constraining condition cannot be replaced with a defining equation such as ( $\downarrow$  TENSE) = PRS because infinitive verbs are assumed not to have the TENSE attribute at all, so the effect of such a defining equation would be to wrongly add the attribute TENSE (and its PRS value) to f-structures of such tenseless forms. Similarly, the constraining equation ( $\uparrow$  SUBJ INDEX PERS) =<sub>c</sub> 3 in the lexical entry (7) for *loves* requires that the f-structure associated with the subject of this verb have the attribute INDEX whose value has the attribute PERS whose value is 3, but the verb does not itself assign this value – some other part of the grammar (in this case, the lexical entry (8) for *she*) must take care of that. As we will see in Section 4.3.3, the existence of such constraining statements presents a difficulty for model-theoretic formalisations of LFG.

While the symbols  $\uparrow$  and  $\downarrow$  only implicitly refer to the function  $\phi$  mapping c-structures to f-structures, the  $\sigma$  function mapping f-structures to s-structures is mentioned explicitly in some of the statements. For example, the statement ( $\uparrow_{\sigma}$  ARG1) = ( $\uparrow$  SUBJ) $_{\sigma}$  in the lexical entry (7) rather concisely says that there is an s-structure associated with the f-structure related to the preterminal V, this s-structure contains the attribute ARG1, and the value of this attribute is the s-structure associated with the f-structure which is the value of SUBJ within the f-structure related to this preterminal. It is easy to check that the representation in Figure 1 reflects this and all the other statements presented in this subsection.

## 3.2 HPSG

Theoretical HPSG grammars have a very different feel: they do not have a CFG backbone, but rather contain statements about various types of linguistic objects – not only phrases and words, but also valencies, contents, cases, etc.<sup>18</sup> HPSG grammars consist of two parts: a type hierarchy (already mentioned in Section 2.2, also called “sort hierarchy” and “signature”) and a theory proper.

A small fragment of the type hierarchy assumed in the AVM of Figure 2 was given in Figure 4, and a much larger part is presented in Figure 7. This type hierarchy seems to mention all types occurring in Figure 2, but in fact it does not contain types for the word forms which appear within PHON values; on the simplest approach to values of PHON each word form is an atom of a type such as

<sup>18</sup>However, some such statements, namely, Immediate Dominance Schemata (Pollard & Sag 1994: Section 1.5), directly encode some of the effects of phrase structure rules.

*she* or *loves*. (On a more comprehensive approach such as Höhle 1999, values of PHON are highly structured and contain various kinds of phonological information.) In a more realistic grammar, the type hierarchy would also contain more subtypes of *headed-phrase* (see Abeillé & Borsley 2021: Sections 5–6 and references therein), a much larger type subhierarchy below *content* (see, e.g., Richter & Sailer 1997, 1999 and Davis 2001 for two very different proposals targeting different aspects of semantic representations), a multiple inheritance hierarchy of subtypes of *head* (Malouf 1998), many more subtypes of *vform*, etc. As shown in Figure 7, type hierarchies are more than just plain taxonomies of types: they also determine attributes that may occur in structures of particular types, as well as types of values of such attributes.

Theory proper is a set of statements – often called principles – which impose additional, more complex constraints. For example, the famous Head Feature Principle (HFP) says that, in a *headed-phrase*, the mother has the same value of the HEAD attribute as the head daughter. Formally, this principle may be stated as follows:

(10) Head Feature Principle:

$$\textit{headed-phrase} \Rightarrow \left[ \begin{array}{l} \text{SYNSEM|CAT|HEAD} \quad \boxed{1} \\ \text{HD-DTR|SYNSEM|CAT|HEAD} \quad \boxed{1} \end{array} \right]$$

Such principles are understood universally: every linguistic object must satisfy them. For this reason they are usually implicational, with the antecedent defining the scope of the principle. In the case of (10), either an object is of type *headed-phrase*, so the antecedent is true and, hence, the consequent must also be true, or the object is not of this type, in which case the antecedent is false and the whole implication is trivially true.

The AVM in Figure 2 describes a configuration of objects containing two objects of type *headed-phrase*, i.e., satisfying the antecedent of HFP: the root object of type *hd-subj-ph* and its HD-DTR value of type *hd-comp-ph*. Both satisfy HFP – see the three occurrences of  $\boxed{10}$  in that figure. All other objects in this configuration satisfy HFP trivially, as they are not described by the antecedent of HFP; this holds for the *word* objects representing *she*, *loves*, and *you*, the *synsem* objects which are values of the SYNSEM attribute, the *list* objects which are values of various occurrences of PHON, SUBJ, COMPS, and DTRS, etc.

There are also constraints relating the values of VAL and DTRS. The role of valency attributes is similar to the role of slashes in Categorical Grammar (Ajdukiewicz 1935; Lambek 1958; see also Kubota 2021 and references therein) – they express information about the combinatory potential of an element. For example, the *word* structure for *loves* in Figure 2 specifies that this word expects

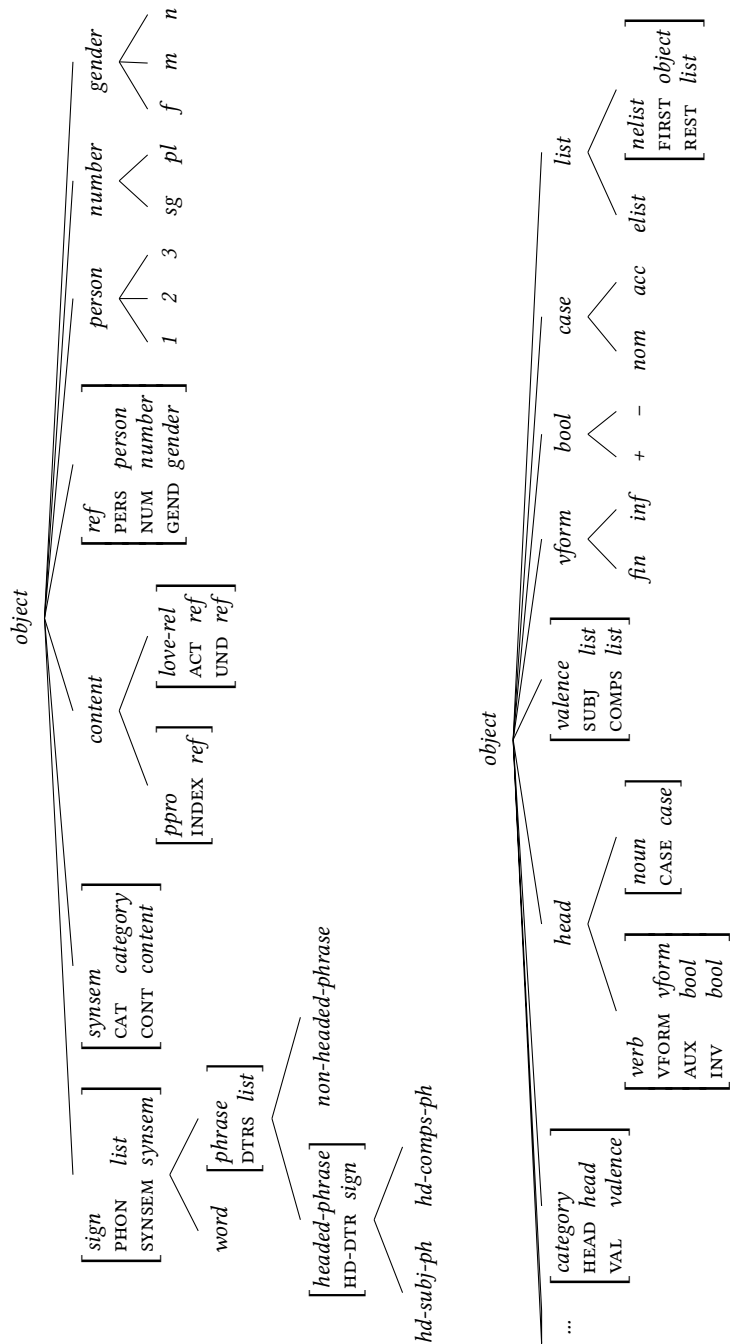


Figure 7: A larger part of an HPSG type hierarchy



a complement and a subject. Once it combines with the complement *you*, the mother phrase of type *hd-comp-ph* needs only a subject in order to be a fully saturated phrase (i.e., a sentence) – its COMPS list is empty (“ $\langle \rangle$ ” is a synonym of the *elist* type in Figure 7). Moreover, once this phrase combines with the subject *she*, both valency lists become empty. This behaviour is regulated by principles such as following:

(11) Valence Principles (modified and simplified):

$$\begin{aligned}
 \text{a. } hd\text{-subj-ph} &\Rightarrow \left[ \begin{array}{l} \text{SYNSEM|CAT|VAL} \left[ \begin{array}{l} \text{SUBJ } \langle \rangle \\ \text{COMPS } \langle \rangle \end{array} \right] \\ \text{DTRS } \langle [\text{SYNSEM } \boxed{2}], [\boxed{1}] \rangle \\ \text{HD-DTR } \boxed{1} \left[ \text{SYNSEM|CAT|VAL} \left[ \begin{array}{l} \text{SUBJ } \langle \boxed{2} \rangle \\ \text{COMPS } \langle \rangle \end{array} \right] \right] \end{array} \right] \\
 \text{b. } hd\text{-comps-ph} &\Rightarrow \left[ \begin{array}{l} \text{SYNSEM|CAT|VAL} \left[ \begin{array}{l} \text{SUBJ } \boxed{0} \\ \text{COMPS } \langle \rangle \end{array} \right] \\ \text{DTRS } \langle \boxed{1}, [\text{SYNSEM } \boxed{2}] \rangle \\ \text{HD-DTR } \boxed{1} \left[ \text{SYNSEM|CAT|VAL} \left[ \begin{array}{l} \text{SUBJ } \boxed{0} \\ \text{COMPS } \langle \boxed{2} \rangle \end{array} \right] \right] \end{array} \right]
 \end{aligned}$$

The constraint (11a) is saying that, in phrases of type *hd-subj-ph*, the head daughter ( $\boxed{1}$ ) only requires a subject (its COMPS list is empty), this subject ( $\boxed{2}$ ) is the (SYNSEM value of the) first daughter of the phrase, while the second daughter ( $\boxed{1}$ ) is the head daughter, and the phrase itself is fully saturated (both SUBJ and COMPS are empty). Similarly, (11b) is saying that, in phrases of type *hd-comps-ph*, the single COMPS element of the head daughter is realised as its second daughter, the first daughter being the head, and the phrase does not expect a complement anymore. On the other hand, it still expects whatever subject (if any) is expected by the head daughter. The actual Valence Principle assumed in HPSG is more general; in particular, it allows for longer COMPS lists and the realisation of multiple arguments in a single local tree (see, e.g., Pollard & Sag 1994: 348).

Note that the values of valency attributes are lists of *synsem* structures (see  $\boxed{2}$  in (11)), not whole phrases. This is an attempt to encode locality constraints on selection: a predicate may specify its arguments only by providing the kind of information that is encoded in SYNSEM values, so it cannot select an argument on the basis of its PHON value or with reference to the internal constituency structure of that argument (as it is encoded in the values of DTRS and HD-DTR).<sup>19</sup>

<sup>19</sup>Note also that these principles do not say anything about values of PHON. We will deal with PHON values shortly, in Section 3.3.1.

Adam Przepiórkowski

What about the lexicon? HPSG has full-fledged theories of the hierarchical lexicon, which make it possible to encode various generalisations across lexical items (see Davis & Koenig 2021 and references therein), but for the purpose of this comparison the simple principle in Figure 8 will do.

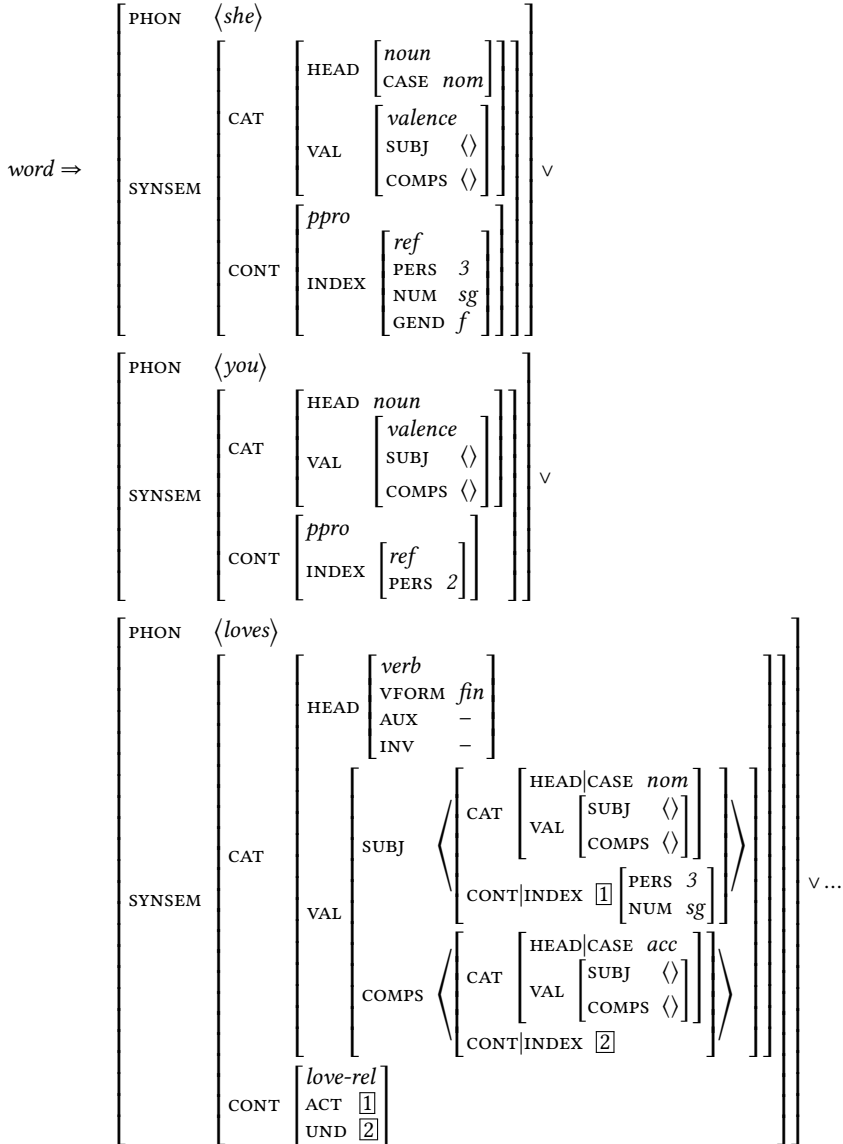


Figure 8: Word Principle

What this principle is saying is that any *word* object must either satisfy the description in the first disjunct (which defines the word *she*), or the second disjunct (*you*), or the third disjunct (*loves*), etc. Again, it is easy to see that the structure described by the AVM in Figure 2 complies with this principle.

All the principles given or alluded to above constrain the shape of *signs* – *words* and *phrases* – but principles may also refer to other types of objects. For example, the type hierarchy in Figure 7 only says that values of SUBJ and COMPS are lists, but the values of SUBJ cannot be of any length – their maximum length is one (a single predicate cannot have two subjects). This can be regulated with the constraint in (12) or – equivalently (given the type hierarchy in Figure 7) but more concisely – (13).

$$(12) \quad \textit{valence} \Rightarrow [\text{SUBJ } \textit{elist}] \vee [\text{SUBJ}|\text{REST } \textit{elist}]$$

$$(13) \quad \textit{valence} \Rightarrow \neg[\text{SUBJ}|\text{REST } \textit{nelist}]$$

Moreover, values of SUBJ and COMPS cannot be just any lists – they must be lists of *synsem* objects. This may be achieved via constraint (14), whose antecedent is not just a type specification, with the predicate *list-of-synsems* defined as in (15).<sup>20</sup>

$$(14) \quad \left[ \begin{array}{c} \text{SUBJ} \quad \boxed{1} \\ \text{COMPS} \quad \boxed{2} \end{array} \right] \Rightarrow \textit{list-of-synsems}(\boxed{1}) \wedge \textit{list-of-synsems}(\boxed{2})$$

$$(15) \quad \textit{list-of-synsems}(\textit{elist}).$$

$$\textit{list-of-synsems}\left(\begin{array}{c} \textit{nelist} \\ \text{FIRST } \textit{synsem} \\ \text{REST } \boxed{0} \end{array}\right) \stackrel{\forall}{\Leftarrow} \textit{list-of-synsems}(\boxed{0}).$$

This simple constraint illustrates an important aspect of contemporary HPSG, namely, the possibility to define and use in constraints any relation (Richter 1999, 2004). The notation for defining such relations is inspired by the programming language Prolog. The definition in (15) consists of two clauses jointly specifying what kinds of objects have the *list-of-synsems* property: the first clause says that the empty list is a list of synsems, and the second (recursive) clause says that a non-empty list whose *FIRST* element is a *synsem* object is a list of synsems if the *REST* of this list is a list of synsems. Nothing else is a list of synsems.

<sup>20</sup>I extend the notational conventions defined in Richter (2004: Section 3.2) in such a way that boxed variables appearing in the antecedents of implications are understood as bound by universal quantifiers scoping over the whole formula. So, the quantificational schema of (14) is:  $\forall \boxed{1} \forall \boxed{2} (\phi(\boxed{1}, \boxed{2}) \Rightarrow \psi(\boxed{1}, \boxed{2}))$ .

3.3 Comparison

3.3.1 Word order

One clear difference between the two frameworks stems from the fact that LFG grammars – but not HPSG grammars – are based on a CFG backbone. Traditionally (but see below) the sentence string is the yield of the c-structure, i.e., it is read off the leaves of the tree. In the case of free word order languages, this leads to trees in which functionally related constituents – for example, a noun and its adjectival modifier – are not always directly related configurationally.

Consider the Warlpiri sentence (16) from Simpson (1991: 257).

- (16) Kurdu-jarra-rlu ka-pala maliki wajili-pi-nyi wita-jarra-rlu.  
child-DU-ERG PRS-3.DU dog.ABS chase-NPST small-DU-ERG  
‘Two small children are chasing the dog.’  
‘Two children are chasing the dog and they are small.’

In this example, *wita-jarra-rlu* ‘small’ is a modifier of *kurdu-jarra-rlu* ‘children’, but on LFG analyses they do not form a constituent, as other constituents linearly intervene between these two words. For example, Austin & Bresnan (1996: 225) propose an analysis which results in the c-structure in Figure 9 (cf. Dalrymple et al. 2019: 112).

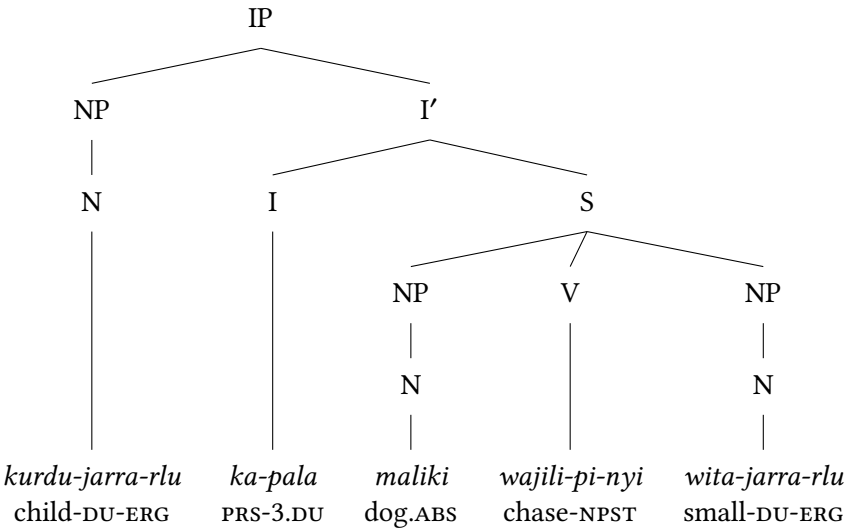


Figure 9: LFG c-structure of (16)

By contrast, it is possible to propose an HPSG analysis of Warlpiri word order on which *wita-jarra-rlu* ‘small’ and *kurdu-jarra-rlu* ‘children’ do form a constituent in (16) (in the sense in which the attribute DTRS represents immediate constituents). A shorthand and very schematic representation of the result of such an analysis is given in Figure 10 (after Donohue & Sag 1999: 13). Note that the order of words within the PHON value of the root of this tree is different from the order of PHON values of the leaves.

This analysis is possible because values of PHON are subject to the same constraints as any other structures. The usual tree behaviour, with PHON values of the mother being the concatenation of the PHON values of the daughters in the order in which they occur on the DTRS list, could be simulated with the constraint in (17), where `append-phones` ( $\boxed{2}, \boxed{1}$ ) holds if  $\boxed{1}$  is the concatenation of PHON values of the elements of  $\boxed{2}$ .<sup>21</sup>

$$(17) \quad \begin{bmatrix} \textit{phrase} \\ \text{PHON } \boxed{1} \\ \text{DTRS } \boxed{2} \end{bmatrix} \Rightarrow \text{append-phones}(\boxed{2}, \boxed{1})$$

If the constraint in (17) were included in the grammar of Warlpiri, the representation in Figure 10 would be ill-formed.

However, other definitions are possible, which relax this usual approach. In fact, there is a long history of such linearisation accounts in HPSG, dating back to Reape (1992, 1996), Kathol & Pollard (1995), and Kathol (1995, 2000) (see also Müller 2021 and references therein); such a relaxed approach to word order is commonly assumed in HPSG analyses of ellipsis and coordination (see Nykiel & Kim 2021, Abeillé & Chaves 2021, and references therein). On such analyses, the two sentences in (18) (from Chaves 2008: 286) have exactly the same constituency structures but differ in PHON values.

- (18) a. Tim gave a rose to Mary and a tulip to Sue.

<sup>21</sup>Formally, the relation `append-phones` is defined as in (i), and the relation `append` it relies on – as in (ii):

- (i) `append-phones` (*elist*, *elist*) .  

$$\text{append-phones} \left( \begin{bmatrix} \text{FIRST} & \boxed{\text{PHON } \boxed{1}} \\ \text{REST} & \boxed{2} \end{bmatrix}, \boxed{3} \right) \stackrel{\vee}{\leftarrow} \text{append-phones}(\boxed{2}, \boxed{4}) \wedge \text{append}(\boxed{1}, \boxed{4}, \boxed{3}) .$$
- (ii) `append` (*elist*,  $\boxed{1}$ *list*,  $\boxed{1}$ ) .  

$$\text{append} \left( \begin{bmatrix} \text{FIRST} & \boxed{1} \\ \text{REST} & \boxed{2} \end{bmatrix}, \boxed{3}$$
*list*,  $\begin{bmatrix} \text{FIRST} & \boxed{1} \\ \text{REST} & \boxed{4} \end{bmatrix} \right) \stackrel{\vee}{\leftarrow} \text{append}(\boxed{2}, \boxed{3}, \boxed{4}) .$

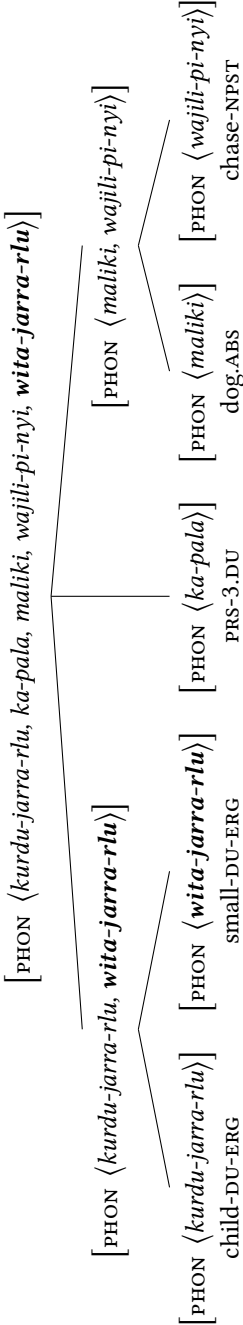


Figure 10: HPSG constituency structure of (16), with the form escaping the default word order constraints marked in bold

- b. Tim gave a rose to Mary and Tim gave a tulip to Sue.

In LFG such a relaxed approach to word order is also in principle possible, on the assumption that there is a representation of the sentence string separate from c-structure. Such a separate string structure – sometimes called s-string (Dalrymple et al. 2019: Section 3.5) – is programmatically proposed in Kaplan (1987) and substantiated in Wescoat (2002), Asudeh (2009) and, especially, Dalrymple & Mycock (2011), among other works, but it is commonly assumed that the order of words in this additional string structure is the same as the order of leaves in c-structure. One exception to this common assumption are the analyses of cliticisation in Bögel et al. (2010) and in Lowe (2016), on which the position of clitics in the s-string may differ from their position in the tree.<sup>22</sup> However, to the best of my knowledge, there are no LFG analyses which would make a more substantial use of the possibility of relaxing the mapping between the s-string and the c-structure, analogous to those common in HPSG accounts of ellipsis.

### 3.3.2 Optionality of attributes

As has already been alluded to above and as will become fully clear in Section 4, grammars may be understood as theories describing certain linguistic objects. Figures such as 1 and 2 are representations of such objects. Both these figures represent all information that follows from all grammatical rules and principles of the respective grammars sketched in this section, but there is a sense in which the LFG representation in Figure 1 is complete while the HPSG representation in Figure 2 is only partial: it represents the effects of all constraints in the grammar proper, but it does not contain all information that follows from the type hierarchy.

Let us have a closer look at the INDEX values within structures corresponding to the word *you*. In both representations in Figures 1 and 2 this value is represented as an AVM with just one attribute, PERS, with a value indicating 2nd person. In the case of LFG, this is a complete description of the underlying feature structure; the linguistic object described by this subsidiary AVM has exactly one attribute: PERS. However, in the case of HPSG, the corresponding AVM is marked as representing a structure of type *ref*(erential index) and – according to the type hierarchy in Figure 7 (and the standard HPSG type system; cf. see Pollard & Sag 1994: 399) – every *ref* object has exactly three attributes: PERS(on), NUM(ber), and GEND(er). Thus, the subsidiary AVM representing the value of INDEX for *you* is

<sup>22</sup>Such a mechanism of prosodic inversion is also alluded to – but not provided an LFG formalisation – in Simpson (1991: 69), Kroeger (1993: 140), and Austin & Bresnan (1996: 226).

*Adam Przepiórkowski*

only a partial description of a complete linguistic object; any such object must also have specific values of NUM (*sg* or *pl*) and GEND (*m*, *f*, or *n*). That is, this subsidiary AVM describes six different kinds of linguistic objects, differing in number and gender.

This technical difference between the two formalisms reflects a potentially important linguistic difference between the two theories: to what extent are the described linguistic objects allowed to be partial or indeterminate?<sup>23</sup> Such partial objects were the staple of the original HPSG of Pollard & Sag (1987), where the described objects were understood not as strictly linguistic objects but rather as informational objects – bits of information (including disjunctive and negative information) that competent speakers have about language.<sup>24</sup> But it seems that LFG sides with the latter-day HPSG in describing linguistic objects rather than informational objects. So the difference between the two representations of the INDEX value for *you* seems to be a linguistically contentful – and potentially verifiable – difference: on the LFG view the pronoun *you* is specified for person but *unspecified* or neutralised for number and gender, while on the HPSG view it is *ambiguous* between different values of number and gender.

Interestingly, it is easy to simulate the HPSG approach in LFG, but it is far from obvious how to simulate the LFG approach in HPSG. In LFG, the lexical entry of *you* could be extended from (9) to (19), with the last two statements requiring that NUM and GEND be present and have values within appropriate sets:

- (19)    *you*                 Pron    ( $\uparrow$  PRED) = ‘PRO’  
                                         ( $\uparrow$  PRONTYPE) = PERSONAL  
                                         ( $\uparrow$  INDEX PERS) = 2  
                                         ( $\uparrow$  INDEX NUM)  $\in \{\text{SG}, \text{PL}\}$   
                                         ( $\uparrow$  INDEX GEND)  $\in \{\text{M}, \text{F}, \text{N}\}$

This leads to six different f-structure representations of the pronoun *you*, differing in the values of NUM and GEND.

Within HPSG, a more complex type subhierarchy could allow for different subtypes of *ref*, one of which would only be specified for the attribute PERS; let us call this subtype *ref-pers*. Then, the pronoun *you* could have the INDEX value of type *ref-pers*, with PERS equal to 2. Another subtype, let us call it *ref-pers-num*, would be specified for PERS and NUM, and it would be appropriate for INDEX values of pronouns *I* (PERS 1, NUM sg) and *we* (PERS 1, NUM pl). This solution, however, is problematic in view of the following examples:

- (20) Creatures, I give you yourselves... (C.S. Lewis, *The Magician's Nephew*)

<sup>23</sup>See also Kaplan (2019) for a discussion of this and related issues.

<sup>24</sup>See Richter (2004: Chapter 2) for a discussion of the differences between early and later HPSG.



(21) Creature, I give you yourself...

The anaphoric pronouns *yourselves* and *yourself* are specified for person (2nd) and for number (plural and singular, respectively), but not for gender, so they should have INDEX values of type *ref-pers-num*. But, given the standard HPSG binding theory (cf. Section 4.3.2 below), these INDEX values should be equal to the INDEX values of the binder – the pronoun *you* in both examples above – so they should be of type *ref-pers*. The only way this is possible is that the two types, *ref-pers* and *ref-pers-num*, have a common subtype. But this common subtype would have to inherit the attribute NUM from *ref-pers-num*, so *ref-pers* would have a subtype with attribute NUM. Given that all objects in HPSG models – including all values of attributes – must bear maximally specific types (this will be made clear in Section 4.1 below), the pronoun *you* would be ambiguous: on one interpretation its INDEX would have a value (of this shared subtype) with NUM sg, on the other – with NUM pl. This would contradict the original motivation for the multiple subtypes of *ref*, namely, to make the pronoun *you* indeterminate with respect to number and gender, rather than ambiguous. It is not clear to me how to simulate within HPSG the behaviour of LFG – that is, how to make the pronoun *you* indeterminate with respect to number by default (i.e., apart from binding contexts such as (20)–(21)) – without complicating the standard HPSG binding theory.

In summary, while either approach may perhaps be simulated in the other theory, HPSG analyses naturally lead to a multiplicity of models differing in ways that linguists often do not care about, while LFG grammars naturally specify fewer linguistic objects, differing only in linguistically relevant aspects. We will return to this issue in Section 4.3.

### 3.3.3 Expressiveness

What is the relation of LFG and HPSG to the Chomsky–Schützenberger hierarchy of grammar formalisms (Chomsky 1956)? That is, what classes of languages do possible LFG and HPSG grammars describe? This question cannot be answered without making the notion of a “possible LFG/HPSG grammar” more precise. Given that both theories evolve and that at any particular point there are competing proposals about various aspects of the theories, this notion is not fully explicit and perhaps never will be.

Nevertheless, it is possible to ask about the complexity of the underlying formalisms, and it is clear that – without additional constraints – both are equivalent to Turing machines, i.e., they may describe any language that is algorithmically

*Adam Przepiórkowski*

describable at all. There is no space here to formally prove this claim, but it is based on the well-known fact that attribute–value grammars may encode Turing machines (Johnson 1988: Section 3.4.2; see also Kaplan & Bresnan 1982: fn. 32). In particular, the unification grammar schema for simulating the effect of any Turing machine (i.e., for defining the same language as that recognised by that Turing machine) presented in Francez & Wintner (2012: Section 6.2) can be easily encoded in the formalisms underlying LFG and HPSG.<sup>25</sup>

Given this formal power of the underlying formalisms, the recognition problem (given a grammar and a sentence, is this sentence predicted by this grammar?) is undecidable – there is no general algorithm which could take an arbitrary grammar and sentence and always answer that question in finite time. In the case of LFG, this potential problem was recognised very early and a solution was proposed (Kaplan & Bresnan 1982: 266–267) in terms of what later became known as offline parsability (Pereira & Warren 1983: 142): a global condition on constituency structures, namely, that, first, they do not contain unary chains (subtrees with only unary branching) in which the same category appears twice and, second, that they do not use empty productions (i.e., that there are no empty leaves in the tree). The encoding of Turing machines in Francez & Wintner (2012: Section 6.2) violates both conditions (cf. fn. 25). A different way to make LFG grammars tractable is proposed – and references to other attempts are given – in Wedekind & Kaplan (2020).<sup>26</sup>

In the case of HPSG, the dominant underlying formalism (RSRL, Richter 1999, 2004; see Section 4.1) is known to be undecidable (Kepser 2004). A different formalisation, based on an extension of modal logic (namely, polyadic dynamic logic), is proposed and shown to have more desirable complexity properties in Søgaaard & Lange (2009) but, to the best of my knowledge, it has remained largely unnoticed within the HPSG community.

---

<sup>25</sup>In the case of LFG, the schemata  $\rho_1$ – $\rho_8$  of Francez & Wintner (2012: 230–232) can be directly translated into LFG phrase structure rules by taking CAT values to be node labels and by encoding all the other information present in the AVMs in  $\rho_1$ – $\rho_8$  via straightforward functional equations. In the case of HPSG, these schemata may be encoded as Immediate Dominance Schemata (Pollard & Sag 1994: Section 1.5), with an additional PHON attribute collecting the terminal symbols (dually to how they are collected in the values of the LEFT attribute in Francez & Wintner 2012). Schemata  $\rho_1$ – $\rho_8$  are essentially – appropriately annotated (which is the source of the additional complexity) – right-linear grammars with binary branching rules for reading the terminal symbols and with unary branching rules – including an empty production – for simulating a Turing machine. It is easy to modify this encoding to get rid of the empty production (the unary rules encoding transitions of a Turing machine could be used at the top of the tree instead of at the right-hand bottom), but the use of effectively unary rules with possible repetitions of non-terminal symbols along unary chains is non-negotiable.

<sup>26</sup>Simplifying, Wedekind & Kaplan (2020) require of grammars that there be an upper bound on the number of different c-structure nodes that may map to the same f-structure.

Let me reiterate, however, that any less than desirable complexity results mentioned above pertain to formalisms underlying the linguistic theories, not to the theories themselves. As has been repeatedly noted in both frameworks (see, e.g., Kaplan & Bresnan 1982: 271–272, Johnson 1988: 94–95, and Richter 2004: 242–243), it is very well possible that linguistic constraints sufficiently delimit the space of possible grammars to make the recognition problem decidable and efficient, and – conversely – it is also possible that human languages are in fact undecidable. That means that high complexity results for a formalism underlying a linguistic theory should not necessarily be held against that theory.

### 3.3.4 Generative-enumerative or model-theoretic?

Pullum & Scholz (2001) divide syntactic frameworks into “generative-enumerative” (GE) and “model-theoretic” (MT). GE frameworks have a derivational feel: at their centre are instructions for rewriting certain strings or structures into other strings or structures. Typical examples are formal grammars in the sense of the Chomsky hierarchy, for example, CFGs such as that in (4'–9'), where particular rules are such instructions. In the top-down mode, one starts with the string “S” and uses the rules to rewrite any non-terminal symbols – e.g., the rule (4') to replace “S” with “NP VP” – until the resulting string contains only terminal symbols, e.g., “*she loves you*”. In the bottom-up mode, one starts with a string of terminal symbols, e.g., “*she loves you*”, and uses the rules in the other direction, until the resulting string “S”, e.g.: “*she loves you*” → “*she loves Pron*” → “*she loves NP*” → “*she V NP*” → “*she VP*” → ... → “S”. The language defined by a grammar is the set of those strings of terminal symbols for which this procedure succeeds. Examples of GE systems are various transformational grammars, Categorical Grammars, Tree-Adjoining Grammars, etc. GE frameworks have an analogue in syntactic – proof-theoretic – aspects of logic.

By contrast, MT frameworks have an analogue in semantic – model-theoretic – aspects of logic. Grammars are sets of logical formulae which may be understood as defining models (namely, those models in which all the formulae are true). An early – historical – example is Arc-Pair Grammar, but currently HPSG seems to be the most clear case of an MT framework (Pullum 2019: 60). We will have a closer look at models of HPSG grammars shortly, in Section 4.1.

Some GE frameworks have a somewhat mixed character: they have a GE backbone but they also impose certain constraints on the resulting structures.<sup>27</sup> Two examples are the transformational grammar of 1980s (GB; Chomsky 1981, 1986)

<sup>27</sup>Thanks to Geoff Pullum for discussion and for the clarification that such “mixed” frameworks should still be classified as unambiguously GE.

*Adam Przepiórkowski*

and, to some extent, Generalized Phrase Structure Grammar (GPSG; Gazdar et al. 1985). It seems that, at least as originally conceived, LFG belongs in the same category: there is a generative CFG backbone responsible for building c-structures (Kaplan & Bresnan 1982: 175), but also for generating functional statements which act as constraints on f-structures associated with particular c-structure nodes (Kaplan & Bresnan 1982: 181). The following quote makes this dual nature of the original LFG particularly clear:

A string's constituent structure is generated by a context-free c-structure grammar. The grammar is augmented so that it also produces a finite collection of statements specifying various properties of the string's f-structure.  
(Kaplan & Bresnan 1982: 180–181)

If such statements – i.e., functional equations – cannot be satisfied, then the whole description for a given input fails, even if the c-structure rules produced an appropriate constituency tree for this input. The functional component thus acts as a filter on the output of the c-structure component (as explicitly stated in Kaplan & Bresnan 1982: 203–204).

Also some general LFG principles are formulated as constraints on possible f-structures (Kaplan & Bresnan 1982: 178–179, Dalrymple et al. 2019: Section 2.4.6): completeness and coherence jointly state that, simplifying a little, grammatical functions mentioned in PRED values must be exactly the grammatical functions occurring as attributes. The main f-structure in Figure 1 satisfies this constraint: PRED mentions SUBJ and OBJ and these are exactly the attributes which characterise grammatical functions in this f-structure. Similarly, f-structures which are values of SUBJ and OBJ satisfy this constraint: their PRED values do not mention any grammatical functions and none appears as an attribute in these f-structures.

Generative-enumerative frameworks may often be given model-theoretic reformulations. McCawley (1968) is usually credited with the observation that phrase structure rules may be understood as conditions on trees,<sup>28</sup> and fully-worked out MT equivalents of various GE formalisms were proposed by Rogers (1997, 1998). While there is no comprehensive MT formalisation of LFG, the description of the general architecture of LFG in Kaplan (1989: Section 2) is formulated in terms of conditions on particular structures, also on c-structures, and on correspondences

---

<sup>28</sup>But cf. Pullum (2007: Section 1.7).

between them,<sup>29</sup> and this view is prevalent in contemporary LFG.<sup>30</sup> For this reason, Pullum & Scholz (2001: 20) classify “recent LFG” as “perhaps” MT. I will have much more to say about model-theoretic aspects of LFG and HPSG in Section 4.

### 3.4 Summary

In this section we looked at two rather specific differences between LFG and HPSG grammars and two more general aspects. One specific difference concerns word order: in HPSG, but not in LFG, the string is often – especially, in analyses of ellipsis – assumed to be dissociated from the constituency structure. The other concerns determinacy: HPSG analyses often lead to multiple structures, i.e., to ambiguity, while LFG analyses more naturally lead to more compact indeterminate structures. Interestingly, despite the expressive power of the two theories, it is not always clear how to elegantly simulate in one theory the analysis commonly assumed in the other.

The two more general issues are expressivity and relation to the generative-enumerative vs. model-theoretic dichotomy postulated in Pullum & Scholz (2001). Underlying formalisms of both theories, unless additionally constrained, have the expressive power of Turing machines; such additional constraints were proposed in LFG right at the beginning and are the topic of ongoing work, while much less attention is devoted to the matter of complexity in HPSG. Finally, HPSG is a prototypical model-theoretic theory, while the place of LFG in this dichotomy is less clear, as no explicit model theory has ever been proposed for LFG. This is the issue to which I turn next.

## 4 Models

Grammars like those discussed in Section 3 are *descriptions* of collections of linguistic objects, pictures like those in Figures 1 and 2 of Section 2 are *representations* of particular configurations of such objects, but what exactly are these *objects* themselves? That is, what are the models of LFG and HPSG grammars?

---

<sup>29</sup>The slightly modified version of Kaplan (1989) published a few years later explicitly invokes “model-based approach” as “of course, the hallmark of LFG” (Kaplan 1995: 11). An earlier model-theoretic formalisation of an LFG-like formalism (but without the distinction between defining and constraining statements) is Johnson (1988). See also Blackburn & Gardent (1995) for another attempt (also limited to defining statements; cf. Börjars & Payne 2013).

<sup>30</sup>For example: “In LFG, phrase structure rules are not rewrite rules, rather they are ‘node admissibility conditions’ (McCawley, 1968); they are constraints rather than procedures.” (Snijders 2015: 61).

Adam Przepiórkowski

The two theories differ considerably in the extent to which answers to these question are provided: fully explicit model theories are proposed in HPSG, but only sketches and intuitive ideas may be found in LFG. For this reason, in this section I start with HPSG. First, however, a few words about models in general.

Take the following formulae of first-order logic:

- (22)  $\forall x. \text{black}(x) \leftrightarrow \neg \text{white}(x)$
- (23)  $\forall x \forall y. \text{bw}(x, y) \rightarrow \text{black}(x) \wedge \text{white}(y)$
- (24)  $\forall x. \text{black}(x) \rightarrow \exists y. \text{white}(y) \wedge \text{bw}(x, y)$
- (25)  $\forall x. \text{white}(x) \rightarrow \exists y. \text{black}(y) \wedge \text{bw}(y, x)$

Together they are saying that everything is either *black* or *white* (see 22) and that there is a relation, *bw*, which holds between *black* things and *white* things (see 23) such that every *black* thing is in this relation with some (at least one) *white* thing (see (24)) and every *white* thing is related to some (at least one) *black* thing (see (25)). Informally speaking, the previous sentence is a description of possible models of formulae (22)–(25). One model is a two-element set such that one element is black, the other is white, and they are related. Another has two black elements and two white elements such that they are pairwise related, i.e., the relation *bw* denotes two pairs of elements. Another – one that also has two black elements and two white elements – is illustrated in Figure 11. The empty set is also a model, and there are infinitely many other models, both finite (of any cardinality apart from 1) and infinite (of any transfinite cardinality).

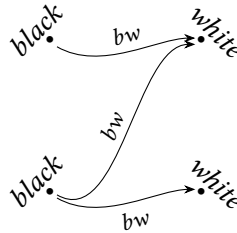


Figure 11: A model – one of many – of (22)–(25)

The meaning of the theory (22)–(25) may be equated with the collection of all models of that theory. However, we may want to exclude some models as not interesting or not really capturing what the formulae (22)–(25) are meant to capture. For example, perhaps we want models to be non-empty and – while possibly arbitrarily large – finite. The first condition may be stated by extending

the theory with the formula  $\exists x. x = x$ .<sup>31</sup> However, the second condition, arbitrary finiteness, cannot be stated within first-order logic, so it must be stated meta-theoretically, as an additional constraint on permitted models.<sup>32</sup> As we will see below, both theories make use of such meta-theoretical conditions on models.

## 4.1 HPSG

Of all linguistic theories, HPSG is perhaps unique in its concentrated attention to the issue of what grammars actually describe – what the models of HPSG theories are. There is no place here to summarise the different proposals found in the HPSG literature; some of them are critically discussed in Richter (2004: Section 2.2). Here, I will describe informally – and in terms which facilitate comparison with standard logical models and with potential LFG models – what I assume to be the standard HPSG approach, namely, the model theory of RSRL (Richter 1999, 2004).<sup>33</sup>

As in mathematical logic, RSRL models are sets of objects which may have various properties and relations defined on them. The properties correspond to the maximal types – called *species* – of type hierarchies: each object is assigned exactly one species.<sup>34</sup> For example, assuming the hierarchy of Figure 7, it is not enough for an object to have the property *list*; it must be either *elist* or *nelist*. Similarly, any *sign* object must actually be either a *word*, or a *hd-subj-ph*, or a *hd-comps-ph*, or a *non-headed-phrase*. In other words, species of HPSG type hierarchies partition sets of objects in HPSG models, just like the properties *black* and *white* partition sets of objects in models of the first-order theory (22)–(25).

Attributes correspond to relations. For example, still assuming the type hierarchy in Figure 7, the attribute *REST* is modelled as a relation between *nelist* objects and *list* (i.e., *elist* and *nelist*) objects. Similarly, *SYNSEM* relates *signs* (i.e., objects of one of the species: *word*, *hd-subj-ph*, *hd-comps-ph*, and *non-headed-phrase*) to objects of type *synsem* (which is a species, according to this type hierarchy). This is similar to the possible interpretations of the relation *bw* as defined in (22)–(25): the domain of that relation is the set of black objects, and the co-domain – the set of white objects. However, the meanings of HPSG attributes are not just any

<sup>31</sup>Given the formulae (22)–(25), the same effect may be achieved, e.g., with  $\exists x. black(x)$  or with  $\exists x. white(x)$ .

<sup>32</sup>Alternatively, a more expressive logic could be adopted.

<sup>33</sup>RSRL – Relational Speciate Re-entrant Language – adds relations and quantification to SRL – Speciate Re-entrant Language – of King 1989, 1999 (see also Pollard 1999). See Richter 2021 for an overview.

<sup>34</sup>That is, in the HPSG lingo, objects are *sort-resolved* (Pollard & Sag 1994: 18).



Adam Przepiórkowski

relations; they are total functions on sets of appropriate species (*nelist*, in the case of *REST*) with values in the set of objects of appropriate species (*elist* and *nelist*, in the case of *REST*).

Additional constraints on objects and relations between them are provided by the theory proper, i.e., by principles such as the HFP in (10), repeated below as (26), and the principles in (11)–(14).

(26) Head Feature Principle:

$$headed\text{-}phrase \Rightarrow \left[ \begin{array}{l} \text{SYNSEM} | \text{CAT} | \text{HEAD} \text{ [1]} \\ \text{HD-DTR} | \text{SYNSEM} | \text{CAT} | \text{HEAD} \text{ [1]} \end{array} \right]$$

For example, HFP is saying that whenever there is an object of type *headedphrase* – i.e., of species *hd-subj-ph* or *hd-comps-ph* – there must be other objects related via functions corresponding to HD-DTR, SYNSEM, CAT, and HEAD as illustrated in Figure 12. In this case, the value of the variable [1] of (26) is the object number 7.

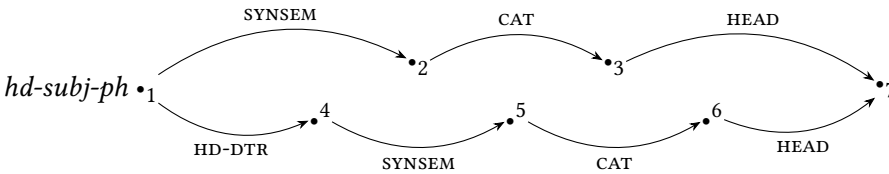


Figure 12: Configuration of objects satisfying the Head Feature Principle

Given the other principles and the type hierarchy, we know much more about this configuration of objects than is explicitly said in Figure 12. For example, the type hierarchy implies that the species of object 4 must be a maximal subtype of *sign*, the species of objects 2 and 5 must be *synsem*, etc. Additional constraints on configurations involving objects of type *hd-subj-ph* are imposed via one of the Valence Principles (namely, (11a)), etc.

Now, HPSG models are simply collections of objects such that each object satisfies all constraints following from the type hierarchy and the theory proper. For example, all seven objects in Figure 12 must satisfy HFP, not just object 1. And they all do, albeit – apart from object 1 – trivially: since objects 2–7 are not of type *headedphrase*, the antecedent of HFP is false of them and the whole statement is true. But the configuration in this figure is not a complete model. For example, according to the type hierarchy in Figure 7, object 7, which is a value of HEAD, must be of type *head*, i.e., of species *verb* or *noun*. If it is a *verb*, there



should be *vform* and *bool* objects in the model related to object 7 via attributes *VFORM*, *AUX*, and *INV*. If it is a *noun*, there should be an object related to object 7 via *CASE*. Similarly, according to the type hierarchy, object 1 should be related to two more objects via *PHON* and *DTRS*, and according to the Valence Principle (11a), the value of *DTRS* should be a two-element list, etc.

Since the late 1980s, all approaches to HPSG models agree with this general view of models, but they all impose additional – technical and sometimes philosophical – constraints on what counts as an interesting model. For example, the empty set is a model (all elements in this set satisfy all constraints), but a trivial one. Also a set consisting of just one object of species *elist* is a model, but it is not interesting. The common view is that HPSG models should be models of whole languages, that they should be exhaustive; in particular, a single exhaustive model contains configurations corresponding to all utterances licensed by the grammar. A little more technically but still very informally, exhaustive models simulate all other models: if there is a structure in some model, then this (or rather, an isomorphic) structure must also occur in an exhaustive model (King 1999). So, within a single exhaustive model, there are configurations of objects corresponding to the AVM in Figure 2,<sup>35</sup> other configurations corresponding to the utterance (2) (*You never give me your money*), and similarly for any other structures licensed by the grammar. This is a somewhat unusual approach to modelling; an analogous exhaustivity requirement in the case of the first-order theory (22)–(25) would mean that only infinite models are admitted, namely those which contain all possible correspondences of black and white objects.<sup>36</sup> We will return to this issue in Section 4.3.

The above considerations still leave open the question: What exactly are the objects in these models? For King (1999) they are bits of reality, actual linguistic tokens (e.g., every utterance of *She loves you* by anybody, ever), but also non-actual – potential – linguistic tokens, i.e., grammatical utterances which have the bad luck of never being actually uttered. This last notion is ontologically dubious, and also leads to proliferation of isomorphic structures within a single model, so it is not frequently subscribed to within the HPSG community.<sup>37</sup> Rather, it is

<sup>35</sup>Recall that the AVM in that figure is still an underspecified description, as it does not fix values of *NUM* and *GEND* within the *ref* object marked as [5]. It is also underspecified in other respects, to be discussed in Section 4.3.2.

<sup>36</sup>In fact, such models would be so large that they would not be sets anymore, but would rather be proper classes.

<sup>37</sup>Also, apart from the curious notion of non-actual tokens, it is not clear what counts as a single utterance token. For example, when John Lennon and Paul McCartney sing together *She loves you*, is this a single token, or two tokens (or perhaps none, because they are singing rather

*Adam Przepiórkowski*

often assumed that the objects in HPSG models are set-theoretic objects – or abstract feature structures – which only stand in conventional correspondence to actual or possible utterances (Pollard & Sag 1994; Pollard 1999). These abstract objects are designed in such a way that – simplifying again – any two isomorphic structures must actually be the same structure. Alternatively, the issue of what exactly these objects are is left unspecified, but an additional requirement is imposed that exhaustive models are minimal in the sense that they contain only one copy of any relevant configuration (Richter 2007).

## 4.2 LFG

While Kaplan (1995: 11) characterises LFG as “model-based”, no explicit and worked-out model theory has ever been proposed for LFG, as far as I know. Let us, nevertheless, try to construct a possible model corresponding to the representation in Figure 1, a model that is consistent with informal descriptions in Kaplan & Bresnan (1982) and Kaplan (1995).

First of all, the model must contain a collection of objects representing the nodes of the c-structure, as well as a collection of node labels (Kaplan 1995: 10). I assume that both grammatical categories (e.g., S or Pron) and orthographic forms (e.g., *she*) are labels. There are three relations defined on these objects: *m* (mother) is the partial function from nodes to nodes, defined on all nodes apart from the root; *<* is the partial ordering relation on nodes, and *λ* is a function from nodes to labels. So a part of the model for the representation in Figure 1, one that corresponds to the c-structure, may look as in Figure 13 (with the linear relation *<* not represented explicitly). There are 18 objects in this model: eight labels (S, VP, V, NP, Pron, *she*, *loves*, *you*) and ten nodes (objects whose exact nature is left unspecified). For an LFG grammar to lead to such models, it must be translated into appropriate formulae, appropriate tree axioms must be stated explicitly, and these axioms should be formulated in such a way that they apply to tree nodes and not to labels or objects corresponding to feature structures.

Kaplan & Bresnan (1982) and Kaplan (1995) are much more explicit about the kinds of objects that correspond to feature structures. There are three types of objects involved in models of feature structures: atoms (e.g., PRED, SUBJ, NOM, 3, etc.), semantic forms (to the first approximation, strings such as ‘LOVE⟨SUBJ,OBJ⟩’), and sets. In particular, feature structures are modelled as finite functions – sets of pairs such that the first element of a pair is an atom and the second element

---

than speaking)? Does the answer depend on whether they sing in unison or in harmony? How many linguistic tokens are there when the song is broadcast on the radio, if any? Does that depend on the number of listeners at different locations?

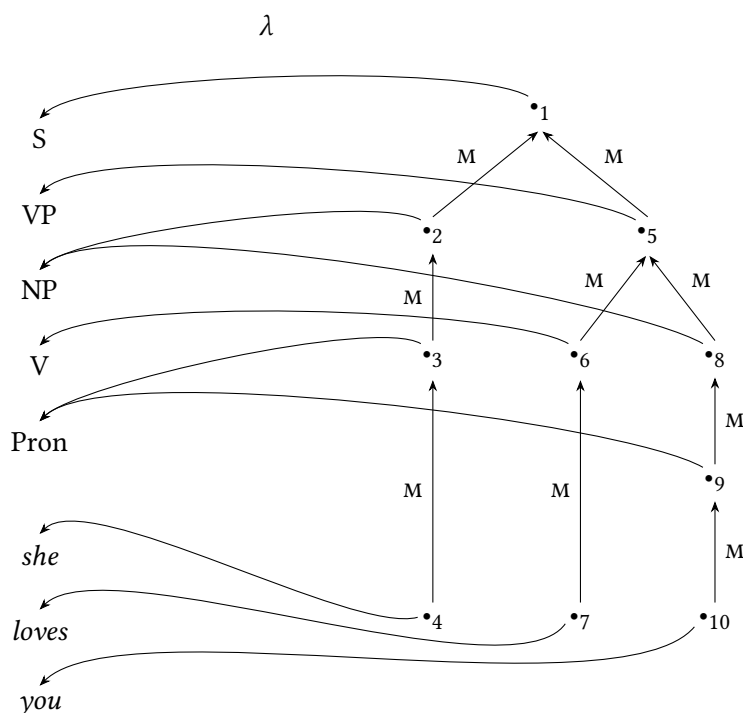


Figure 13: A possible LFG model of the c-structure of *She loves you* (without explicit representation of <)

is either an atom, or a semantic form, or a feature structure (i.e., a set again).<sup>38</sup> For example, the AVM in (27) is a representation of the set of pairs in (28), i.e. – given the commonly assumed Kuratowski's encoding of a pair  $\langle a, b \rangle$  as the set  $\{\{a\}, \{a, b\}\}$  (see, e.g., Enderton 1977: 36) – the set in (29).<sup>39</sup>

<sup>38</sup>Together with Kaplan & Bresnan (1982) and Kaplan (1995), I do not take into consideration sets other than those which model feature structures, i.e., I ignore here coordinate structures, values of the attribute ADJUNCT, etc.

<sup>39</sup>One potential problem with this standard LFG understanding of f-structures as sets is that, given the possibility of cyclic f-structures – naturally occurring in analyses of various types of modification, e.g., Johnson (1988: 19–20), Zweigenbaum (1988), and Haug & Nikitina (2012: 298), and in other contexts, e.g., Fang & Sells (2007: 209), Przepiórkowski & Patejuk (2012: Section 4.3.2), and Dalrymple et al. (2020) – sets that are used for modelling f-structures are not the well-founded sets of the standard (Zermelo–Fraenkel) set theory, but must rather rely on the non-standard notion of non-well-founded sets (see Aczel 1988: 103–112 on the history of this notion). To the best of my knowledge, this has not been noticed in the LFG literature so far.

Adam Przepiórkowski

$$(27) \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{sg} \\ \text{GEND} & f \end{bmatrix}$$

$$(28) \{ \langle \text{PERS}, 3 \rangle, \langle \text{NUM}, \text{SG} \rangle, \langle \text{GEND}, F \rangle \}$$

$$(29) \{ \{ \text{PERS} \}, \{ \text{PERS}, 3 \}, \{ \{ \text{NUM} \}, \{ \text{NUM}, \text{SG} \} \}, \{ \{ \text{GEND} \}, \{ \text{GEND}, F \} \} \}$$

Since some parts of f-structures (values of particular attributes, as well as attributes themselves) may be directly referred to in functional equations, they must all be direct elements of the model. That is, sets representing f-structures cannot be considered unanalysable elements of models; rather, the subsets and atoms within such sets must also be elements of LFG models, so they should be explicitly related by the (converse of the) membership relation  $\in$ . Hence, the set in (29) corresponding to the f-structure (27) translates into the configuration of model objects in Figure 14. There are 10 nodes in this configuration that encode particular sets (with node 1 representing the whole f-structure (27)) and six nodes are atoms.

The model in Figure 14 is rather complex, when compared to the simplicity of the AVM in (27). Why not assume the model in Figure 15 instead?<sup>40</sup> Unfortunately, as explained in more detail presently (in Section 4.3.1), this simpler model is incompatible with the LFG idea that attributes and atomic values are ontologically the same kinds of entities, namely, atoms. By contrast, according to the model in Figure 15, atomic values are atoms – objects of the universe of the model – but attributes are binary relations on such objects, i.e., ontologically very different entities. Hence, in the following I will assume the model in Figure 14 as most directly reflecting the LFG view that f-structures are finite functions.

Let me finish this section by noting that configurations in Figures 13 and 14 are fragments of a larger model corresponding to the representation of *She loves you* given in Figure 1. The complete model would also contain strings representing semantic forms, as well as more atoms, many more sets representing the full f-structure, sets representing the s-structure, and relations  $\phi$  and  $\sigma$ .

## 4.3 Comparison

### 4.3.1 Modelling feature structures

It should be clear from the above discussion that AVM representations correspond to very different model configurations in the two theories. For example,

<sup>40</sup>Compare the HPSG model of (30) in Figure 16 below. Such simpler models, in which feature structures are represented as objects and attributes as relations on objects, are also common in other theories working with AVMs (see, e.g., Blackburn & Spaan 1993: 132–133).

while the HPSG model of the AVM in (30), shown in Figure 16, contains just four nodes corresponding directly to the whole index (object 1 of species *ref*), to 3rd person (object 2 of species 3), to singular number (3 – *sg*), and to feminine gender (4 – *f*), the LFG model of the corresponding AVM in (27), shown in Figure 14, contains 16 nodes modelling not only the whole f-structure and the respective values of the three attributes, but also the attributes themselves and various intermediate sets.

$$(30) \quad \left[ \begin{array}{l} \textit{ref} \\ \text{PERS} \quad 3 \\ \text{NUM} \quad \textit{sg} \\ \text{GEND} \quad \textit{f} \end{array} \right]$$

This is not an incidental difference between the two theories. In HPSG, attributes such as PERS, NUM, and GEND and types such as 3, *sg*, and *f* have very different interpretations: attributes denote relations (partial functions) between objects in the model, while types denote properties that objects may have. In particular, different objects may – and often do – have the same species, so there can be many objects of type *sg*, etc. This difference between attributes and types is rendered typographically by using small capitals for attributes and italics for types.

On the other hand, in LFG, attributes such as PERS, NUM, and GEND and their atomic values such as 3, *sg*, and *f* are the same kinds of objects, namely atoms, each of which may occur in the model just once (there is only one atom *sg*, etc.). Hence, there is also no typographic distinction between attributes and atomic values of attributes.

This ontological uniformity of attributes and atomic values is taken advantage of in some LFG analyses. For instance, according to the analysis of oblique arguments in Kaplan & Bresnan (1982: 196–201), a “case-marking” preposition which may introduce such an oblique argument defines the value of the attribute PCASE to be the oblique function homonymous with this preposition, e.g.:

$$(31) \quad \textit{to} \quad \text{P} \quad (\uparrow \text{PCASE}) = \textit{TO}$$

This feature and its value are also present in the f-structure corresponding to the resulting PP constituent. Verb forms like *handed*, as used in (32) from Kaplan & Bresnan (1982: 196), expect – apart from any subject and objects – an argument bearing this grammatical function, see (33).

$$(32) \quad \text{A girl handed a toy to the baby.}$$

Adam Przepiórkowski

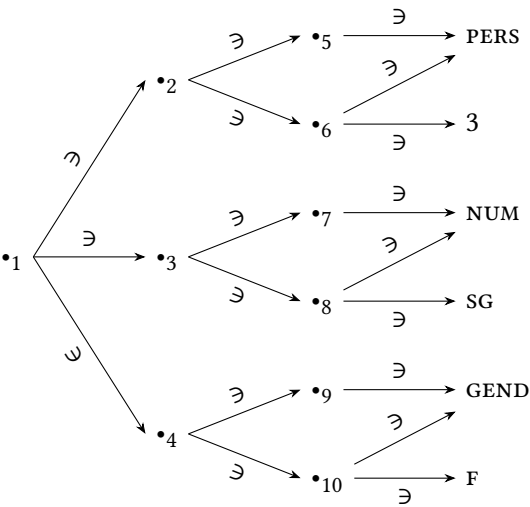


Figure 14: A possible LFG model of the f-structure in (27)

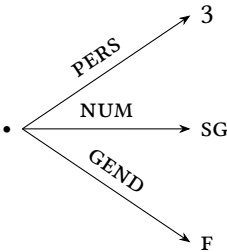


Figure 15: A hypothetical simpler model of the AVM in (27)

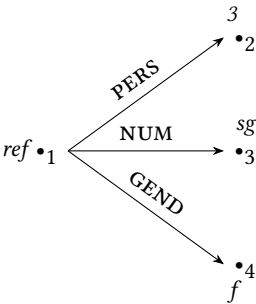


Figure 16: An HPSG model of the AVM in (30)

$$(33) \quad \textit{handed} \quad V \quad (\uparrow \text{ PRED}) = \text{'HAND<SUBJ,OBJ,TO>'} \\ (\uparrow \text{ TENSE}) = \text{PST}$$

Finally, an appropriate VP rule – simplified here to (34) – contains the crucial equation (35) on the PP:

$$(34) \quad VP \longrightarrow \quad V \quad \quad \quad NP \quad \quad \quad PP \\ \downarrow = \uparrow \quad (\uparrow \text{ OBJ}) = \downarrow \quad (\uparrow (\downarrow \text{ PCASE})) = \downarrow \\ (\downarrow \text{ CASE}) = \text{ACC}$$

$$(35) \quad (\uparrow (\downarrow \text{ PCASE})) = \downarrow$$

Applied to the sentence (32), with the PP *to the baby*,  $(\downarrow \text{ PCASE})$  in equation (35) evaluates to *to*, so the whole equation is equivalent to  $(\uparrow \text{ to}) = \downarrow$ . Note that *to*, the atomic value of *PCASE* of the preposition *to*, is used here as an attribute indicating an oblique grammatical function. While such double use of atoms as values and attributes is rare in actual LFG analyses, it is not unique to the account of obliques in Kaplan & Bresnan (1982); for example, it also occurs in the formalisation of information structure in Dalrymple & Nikolaeva (2011: Sections 4.3.3–4.3.5).

The above considerations do not imply that not distinguishing attributes from atomic values necessarily leads to such complex models as that partially illustrated in Figure 14. For example, Johnson (1988: Section 2.1.3) defines models of f-structures as consisting of a set of atoms, a set of objects directly modelling particular feature structures, and a 2-argument partial function  $\delta$  whose first argument is an f-structure, second argument is an atom *qua* attribute, and the value is the value of this attribute in this f-structure. On this approach the AVM in (27) receives a model that may be represented pictorially as in Figure 17. Note, however, that on this view feature structures are no longer sets of  $\langle$  attribute, value  $\rangle$  pairs, contrary to Kaplan & Bresnan (1982) and Kaplan (1995).

#### 4.3.2 Identity of indiscernibles?

Both theories have trouble with indiscernible structures. Let us illustrate this with sentence (36).

(36) She says she loves you.

Consider the LFG f-structure for this sentence in Figure 18. In the model configuration corresponding to this AVM there are single objects representing particular atoms: just one object *NOM*, one *SG*, one *PRS*, one *TENSE*, etc. Moreover, since feature structures are sets of  $\langle$  attribute, value  $\rangle$  pairs, the two *INDEX* values – the substructures marked as [2] and [4] – are the same set (namely, the one in

Adam Przepiórkowski

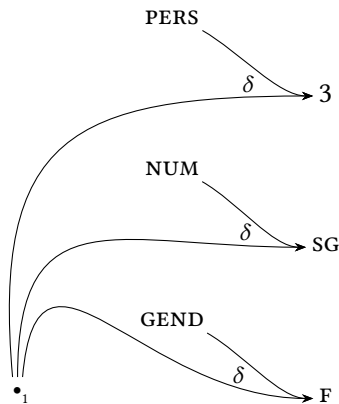


Figure 17: A model of the f-structure in (27) as in Johnson (1988)

(29)), so they should be modelled with the same single object in the model (or, more precisely, with a single configuration of objects shown in Figure 14, rooted in the same object 1). The problem is that nothing in our reconstruction of the intended LFG model theory guarantees this: two different models are possible, one in which  $[2] = [4]$ , and one in which  $[2] \neq [4]$ . Only the first of these models properly encodes the idea that feature structures are sets.<sup>41</sup> We will return to this issue below, when discussing HPSG models.

The bigger problem is that, if f-structures are sets, the two f-structures representing *she*, i.e.,  $[1]$  and  $[3]$  in Figure 18, are the same set-theoretical object. (In the modelling of f-structures suggested above they may be the same object, but – as discussed in the previous paragraph – they do not have to be.) But LFG requires that they be different objects – we do not want to say that the two ‘PRO’ values in these f-structures necessarily refer to the same person. The way LFG deals with this problem is to assume that PRED values – semantic forms – come with unique indices (normally not shown in AVMs), i.e., that whenever an equation like  $(\uparrow \text{PRED}) = \text{‘PRO’}$  is used, a new index is assigned to the semantic form. So the two references to the lexical entry for *she* in (8) that are made in the process of constructing the f-structure in Figure 18 result in two different equations, as if the following two statements were used:

(37) a.  $(\uparrow \text{PRED}) = \text{‘PRO’}_1$

<sup>41</sup>Interestingly, the XLE platform for implementing LFG grammars (Crouch et al. 2008), normally very faithful to the LFG theory, does not treat f-structures as (standard) sets: there, two indiscernible f-structures are assumed to be different objects, unless there is a statement in the grammar that explicitly requires their identity.



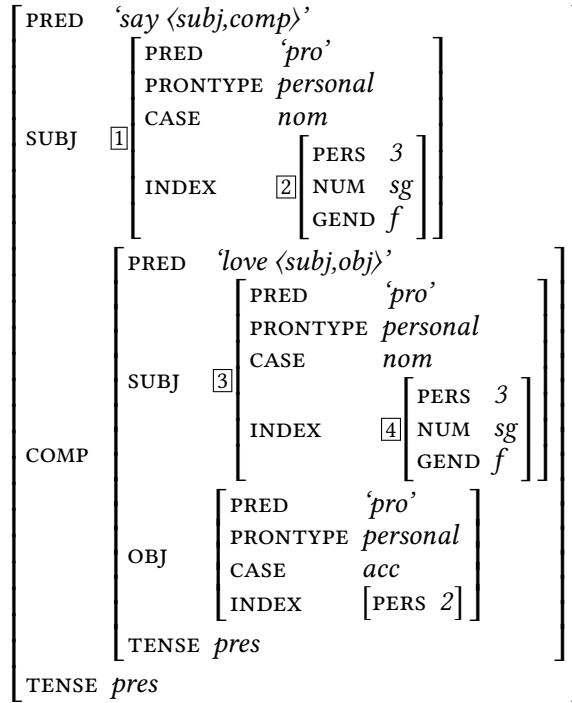


Figure 18: F-structure for (36)

b. ( $\uparrow$  PRED) = ‘PRO’<sub>2</sub>

Unfortunately, this mechanism, as it stands, seems to be inherently procedural: at the relevant step of the derivation it must be known which indices have already been used so that a new index can be assigned to a new semantic form. It is not immediately clear how to translate this mechanism to the model-theoretic view of LFG.<sup>42</sup>

Also HPSG has a problem with stating when exactly indiscernible structures should be treated as being the same structure.<sup>43</sup> In HPSG, not even atoms are

<sup>42</sup>Given that PRED values are largely redundant (cf. Section 2.1 and fn. 12), this problem may be solved by removing PRED from f-structures altogether. Another – perhaps more conservative – possible solution, suggested by Ash Asudeh (p.c.), is to provide indices with sufficient inherent structure to guarantee their uniqueness. In the simple case of (36), it would suffice to take indices to be the relevant c-structure nodes, but a more complex solution is required to also apply to ‘pro’ values of PRED in the case of pro-dropped constituents (especially in languages which allow pro-dropping of multiple arguments of a single predicate).

<sup>43</sup>The problem to be described presently is sometimes called “Höhle’s problem” (Pollard 2001, 2014: 113).

Adam Przepiórkowski

guaranteed to be unique, so one of the models of sentence (36) (*She says she loves you*), whose partial AVM is given in Figure 19, might involve the configuration in Figure 20, with single objects of type 3 and *sg* and two different objects of type *f*. Given two different *ref* objects, there are eight possible configurations of this part of the model, and given also the possibility of two different *nom* objects, two different *she* objects (in *PHON* values), different *elist* objects, etc., there are billions of different models of sentence (36), all described by the AVM in Figure 19, differing in ways that linguists do not care about.<sup>44</sup> This contrasts with the efforts within HPSG (Pollard & Sag 1994; Pollard 1999; Richter 2007) to make models of various interpretations of utterances unique (at least up to isomorphism).

Now, it is possible to formulate within RSRL a constraint that makes sure that all indiscernible structures are in fact the same structure (Sailer 2003: Section 3.1.4), but such a constraint, if applied indeterminately, would be incompatible with various HPSG analyses – most importantly, with the standard HPSG binding theory (Pollard & Sag 1994: Chapter 6).<sup>45</sup> There is no space to present that theory here (see Müller & Branco 2021 for an overview), but suffice it to say that the traditional generative notion of coindexation is understood here literally: as identity of *INDEX* values. For example, the sentence (36) is assumed in HPSG to have two different structures corresponding to the following two indexations:

- (38) a.  $She_i$  says  $she_i$  loves you.  
       b.  $She_i$  says  $she_j$  loves you. ( $i \neq j$ )

So while any model of (38a) should equate *INDEX* values within the two words *she* in this sentence, these *INDEX* values must be different objects in any model of (38b), even though they are indiscernible.

To the best of my knowledge, the problem of avoiding spuriously distinct models in a way that does not conflict with existing HPSG analyses (in particular, with the standard binding theory) remains unsolved.

<sup>44</sup>Each word introduces three lists (values of *PHON*, *VAL|SUBJ*, and *VAL|COMPS*), and there are five words in this sentence, so there are 15 *elist* objects stemming from words alone. The number of different ways to partition a set of *n* elements into equivalence classes is given by Bell numbers  $B_n$ , and  $B_{15} = 1,382,958,545$  (see <https://oeis.org/A000110/list>). This should be multiplied by the eight configurations of the two *ref* objects, etc. Richter (2007) proposes a constraint to the effect that all *elist* objects are the same object, but the problem of the other spurious ambiguities remains.

<sup>45</sup>Also the architecture for phonology proposed in Höhle (1999) crucially relies on not all indiscernible structures being the same structure. Sailer (2003) formulates the relevant constraint in such a way that it only applies to one type of structures.

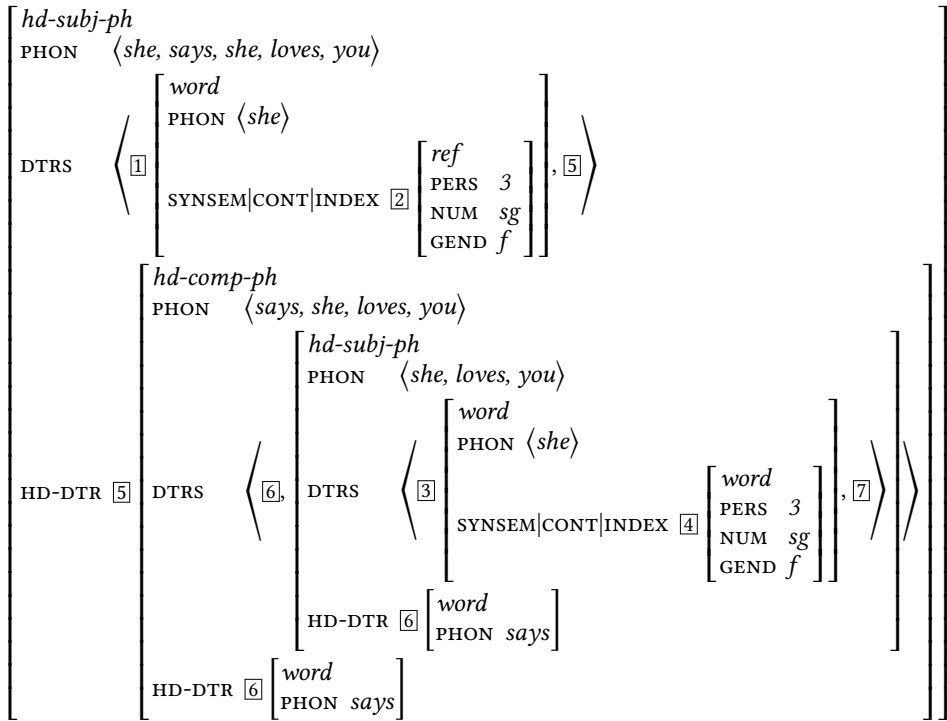


Figure 19: Partial HPSG representation of (36)

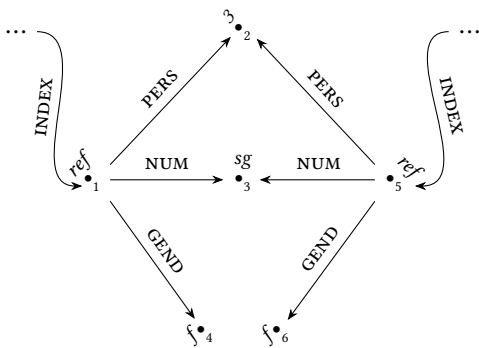


Figure 20: A fragment of a possible HPSG model of the AVM in Figure 19

*Adam Przepiórkowski*

#### 4.3.3 Conditions on models

Both theories impose meta-theoretical conditions on what counts as an intended model. As mentioned in Section 4.1, the common constraint on HPSG models is that they be exhaustive, i.e., informally speaking, simulate all other models: they should contain all structures admitted by the grammar. The intuition behind this requirement is that a single model corresponds to the whole language described by the grammar.

LFG apparently assumes the more common view of models, where each model corresponds to a single utterance, and it is only the collection of all such models that corresponds to the whole language. However, meta-theoretical conditions on LFG models are in a way more complex than conditions imposed on HPSG models.

First of all, LFG models are required to be minimal. For example, functional equations in the lexical entry of *she* (see (8)) involving the attribute INDEX, i.e. equations repeated below in (39), describe as a possible value of INDEX not only the feature structure in (27), repeated below as (40), but also the one in (41) and infinitely many others, including infinite feature structures (both infinitely embedded and – on the assumption that the set of atoms may be infinite – with an infinite number of attributes).

$$(39) \quad \begin{aligned} (\uparrow \text{ INDEX PERS}) &= 3 \\ (\uparrow \text{ INDEX NUM}) &= \text{SG} \\ (\uparrow \text{ INDEX GEND}) &= \text{F} \end{aligned}$$

$$(40) \quad \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{sg} \\ \text{GEND} & f \end{bmatrix}$$

$$(41) \quad \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{sg} \\ \text{GEND} & f \\ \text{ARBI} & \text{trary} \\ \text{NON} & [\text{SEN } se] \end{bmatrix}$$

Other constraints in the grammar do not preclude such values of INDEX, so a meta-theoretical constraint is needed to the effect that only minimal feature structures satisfying the grammar are admitted within models. Technically, this amounts to defining a partial order on models and admitting only the minimal elements of this order.

The second condition on models is more complex and concerns constraining statements such as (42a) (from the grammar rule (4)) and (42b) (from the lexical entry (7)).

- (42) a. ( $\downarrow$  TENSE)  
 b. ( $\uparrow$  SUBJ INDEX PERS) =<sub>c</sub> 3

Such statements are understood as additional filters on the minimal models of a grammar, or – more precisely – on the minimal models of the version of the grammar with all such constraining statements removed.

The precise model-theoretic nature of this mechanism has never, to the best of my knowledge, been specified. Constraining statements of this kind are not mentioned in the model-theoretic view of LFG of Kaplan (1995), and they are explicitly excluded in previous attempts to provide LFG (or LFG-like) formalisms with a model theory (see Johnson 1988: Section 4.2 and Blackburn & Gardent 1995: Section 6; see also Börjars & Payne 2013). But once meta-theoretical quantification over models and relations on models are permitted – and they are already inherent both in the HPSG notion of exhaustive models and the LFG notion of minimal models – it is possible to understand constraining statements in model-theoretic terms. One possibility is this:<sup>46</sup>

- Let  $\theta$  be an LFG grammar, understood as a set of logical formulae. Some of the (sub)formulae are marked as constraining, the others are understood as defining.
- Let  $\theta_{all}$  be the whole grammar  $\theta$  without any division of (sub)formulae into defining and constraining, and  $\theta_{def}$  – the same grammar with all constraining (sub)formulae removed.
- Let  $M_{all}$  be the collection of all models of  $\theta_{all}$ , and  $M_{def}$  – the collection of all *minimal* models of  $\theta_{def}$ .<sup>47</sup>
- Then  $M \stackrel{\text{df}}{=} M_{def} \cap M_{all}$  is the collection of admitted models of  $\theta$ .

<sup>46</sup>Given that statements may contain disjunctions, and that different constraining statements may occur in different disjuncts, the actual definition would have to be more complex: grammars would have to be converted to a disjunctive normal form and collections of models would have to be defined for each disjunct of this normal form. Then the final collection of models of the grammar would be the sum of all such collections.

<sup>47</sup>Formally, minimal models are the minimal elements of the subsumption relation defined on models as in Johnson (1988: Section 2.8).

*Adam Przepiórkowski*

The idea here is that  $M_{def}$  is the collection of all minimal models before the constraining filters are applied, and the intersection with  $M_{all}$ , i.e., with models in which all constraining statements are satisfied, removes from  $M_{def}$  those models which do not satisfy some constraining statements.

#### 4.4 Summary

This section, aiming to present and compare model theories assumed in HPSG and LFG, is more speculative than the previous sections. The reason is that one object of comparison exists and the other does not, so it was necessary to reconstruct a possible model theory of LFG from informal and very partial suggestions.

Perhaps surprisingly, it turns out that the idea that f-structures are sets of  $\langle \text{attribute}, \text{value} \rangle$  pairs does not translate into elegant models, but rather creates an overhead of the need to represent these sets as objects within models. Also, additional care needs to be taken to ensure that co-extensional sets are really the same model objects. Moreover, it is not immediately clear how to formally and non-procedurally ensure unique indexation of semantic forms. Nevertheless, despite these difficulties, and despite the fact that constraining statements were excluded from previous attempts to construct a model theory for LFG, it is not difficult to imagine how to construct such a model theory, if only appropriately powerful meta-theoretical operations on candidate models are permitted (as – to some extent – they already are, given the minimality requirement).

Also somewhat surprisingly, while much attention has been devoted to model theory within HPSG, there are still unsolved problems there, concerning the multiplicity of different models admitted by typical HPSG grammars, differing in ways that linguists often do not suspect, and certainly do not care about.

Let the conclusion of this section be that, despite their age and stability, both theories would benefit from more work on their formal foundations.

### 5 Conclusion

So how similar are LFG and HPSG? I agree with Carl Pollard that in some ways they are more similar than sometimes perceived:

I believe that the difference between LFG and so-called PSG [i.e., theories such as GPSG and HPSG; AP] is no greater than the differences among various theoretical proposals within PSG, or even within HPSG itself. As far as I am concerned, then, the separation between PSG and LFG exists more at

a sociological level than at the level of scientific content – but I am aware that not everyone agrees about this. (Pollard 1997: 4)

In particular, the difference between the multi-level representations of LFG and the monolithic AVMs assumed in HPSG is – in my view – of little formal consequence, although it is certainly important for the compactness and readability of resulting structures.

In fact, LFG and HPSG converge in many respects. As emphasised above, both theories are highly formalised and – unlike derivational theories or Categorical Grammar – both are self-described as constraint-based or model-theoretic, although HPSG may boast of much more developed model theories. Importantly, both have well-developed computational platforms for implementing grammars: XLE (Crouch et al. 2008) in the case of LFG and LKB (Copestake 2002) and Trale (Carpenter et al. 2003) in the case of HPSG, with XLE allowing for very direct implementations of theoretical analyses.<sup>48</sup> In both cases, large-scale grammars of multiple languages have been developed.

Also, unlike some of the other highly formalised and implementable theories, both LFG and HPSG are empirically rich. A plethora of analyses of multiple phenomena in typologically varied languages have been offered within each theory, in a great many articles appearing in the best linguistic journals and in numerous monographs published by the most prominent publishers. Both have very well developed semantic components, and both make it possible to formulate precise analyses encompassing multiple linguistic levels. As emphasised in Wechsler & Asudeh (2021), many phenomena receive similar accounts in the two theories.

In summary, it is clear that LFG and HPSG are close neighbours in the linguistic theoretical landscape of the early 2020s, and it is my hope that this chapter encourages more neighbourly collaboration between the two theories.

## Acknowledgements

Many thanks to the following people for comments on the first version of this chapter and – in some cases – for discussion: Doug Arnold, Ash Asudeh, Stefan Müller, Agnieszka Patejuk, Geoff Pullum, Frank Richter, Shuly Wintner, Annie Zaenen, and the anonymous reviewers. They should not be held responsible

<sup>48</sup>More precisely, XLE makes it possible to faithfully implement syntactic and – with extensions described in Dalrymple et al. (2020) – semantic parts of LFG analyses. In the case of HPSG, Trale seems to be closer to its constraint-based nature, while LKB is more efficient and well-developed.

Adam Przepiórkowski

for any controversial opinions expressed here or for any remaining inadequacies; if some of their suggestions are less than fully addressed, it is only because this is just a chapter (and one that is too long already!), and not the monograph that a fuller presentation and comparison of LFG and HPSG deserves. And many thanks to Mary Dalrymple, the editor of this volume, for her infinite patience!

## References

- Abeillé, Anne & Robert D. Borsley. 2021. Basic properties and elements. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Abeillé, Anne & Rui P. Chaves. 2021. Coordination. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Ackerman, Farrell & Gert Webelhuth. 1998. *A theory of predicates*. Vol. 76 (CSLI Lecture Notes). Stanford, CA: CSLI Publications.
- Aczel, Peter. 1988. *Non-well-founded sets*. Stanford, CA: CSLI Publications.
- Ajdukiewicz, Kazimierz. 1935. Die syntaktische Konnexität. *Studia Philosophica* 1. 1–27.
- Alsina, Alex. 1996. *The role of argument structure in grammar: Evidence from Romance* (CSLI Lecture Notes). Stanford, CA: CSLI Publications.
- Arnold, Doug & Danièle Godard. 2021. Relative clauses in HPSG. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Asudeh, Ash. 2009. Adjacency and locality: A constraint-based analysis of complementizer-adjacent extraction. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 106–126. Stanford, CA: CSLI Publications.
- Asudeh, Ash & Gianluca Giorgolo. 2012. Flexible composition for optional and derived arguments. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 64–84. Stanford, CA: CSLI Publications.
- Austin, Peter K. & Joan Bresnan. 1996. Non-configurationality in Australian aboriginal languages. *Natural Language & Linguistic Theory* 14. 215–268. DOI: 10.1007/bf00133684.
- Bird, Steven & Ewan Klein. 1994. Phonological analysis in typed feature systems. *Computational Linguistics* 20(3). 455–491.



- Blackburn, Patrick & Claire Gardent. 1995. A specification language for Lexical-Functional Grammars. In *Proceedings of the 7th conference of the European chapter of the ACL (EACL 1995)*, 39–44. European Association for Computational Linguistics. DOI: 10.3115/976973.976980.
- Blackburn, Patrick & Edith Spaan. 1993. A modal perspective on the computational complexity of attribute value grammar. *Journal of Logic, Language and Information* 2. 129–169. DOI: 10.1007/bf01050635.
- Bögel, Tina. 2021. Prosody and its interfaces. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*, 95–137. Berlin: Language Science Press. DOI: ??.
- Bögel, Tina, Miriam Butt, Ronald M. Kaplan, Tracy Holloway King & John T. Maxwell III. 2010. Second position and the prosody-syntax interface. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 106–126. Stanford, CA: CSLI Publications.
- Börjars, Kersti & John Payne. 2013. Dimensions of variation in the expression of functional features: Modelling definiteness in LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '13 conference*, 152–170. Stanford, CA: CSLI Publications.
- Borsley, Robert D. & Berthold Crysmann. 2021. Unbounded dependencies. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Bresnan, Joan, Ash Asudeh, Ida Toivonen & Stephen Wechsler. 2016. *Lexical-Functional Syntax*. 2nd edn. (Blackwell Textbooks in Linguistics 16). Malden, MA: Wiley-Blackwell.
- Butt, Miriam, Mary Dalrymple & Anette Frank. 1997. An architecture for linking theory in LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '97 conference*, 1–16. Stanford, CA: CSLI Publications.
- Carpenter, Bob, Gerald Penn & Mohammad Haji-Abdolhosseini. 2003. *The Attribute Logic Engine User's Guide with TRALE Extensions*. Version 4.0 Beta. [http://www.ale.cs.toronto.edu/docs/man/ale\\_trale\\_manual.pdf](http://www.ale.cs.toronto.edu/docs/man/ale_trale_manual.pdf).
- Carpenter, Robert L. 1992. *The logic of typed feature structures*. Cambridge. UK: Cambridge University Press. DOI: 10.1017/cbo9780511530098.
- Chaves, Rui P. 2008. Linearization-based word-part ellipsis. *Linguistics and Philosophy* 31. 261–307. DOI: 10.1007/s10988-008-9040-3.
- Chaves, Rui P. 2021. Island phenomena and related matters. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.

Adam Przepiórkowski

- Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2(3). 113–124. DOI: 10.1109/tit.1956.1056813.
- Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris Publications. DOI: 10.1515/9783110884166.
- Chomsky, Noam. 1986. *Barriers*. Cambridge, MA: The MIT Press.
- Copestake, Ann. 2002. *Implementing typed feature structure grammars*. Stanford, CA: CSLI Publications.
- Crouch, Richard, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III & Paula Newman. 2008. *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. 2017. Unlike phrase structure category coordination. In Victoria Rosén & Koenraad De Smedt (eds.), *The very model of a modern linguist – In honor of Helge Dyvik*, vol. 8, 33–55. Bergen: Bergen Language & Linguistics Studies (BeLLS). DOI: 10.15845/bells.v8i1.1332.
- Dalrymple, Mary, Angie Hinrichs, John Lamping & Vijay Saraswat. 1993. The resource logic of complex predicate interpretation. In *Proceedings of ROCLING 1993*, 3–21. <http://www.aclclp.org.tw/rocling/1993/K01.pdf>.
- Dalrymple, Mary, Ronald M. Kaplan & Tracy Holloway King. 2004. Linguistic generalizations over descriptions. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*, 199–208. Stanford, CA: CSLI Publications.
- Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III & Annie Zaenen (eds.). 1995. *Formal issues in Lexical-Functional Grammar*. Stanford, CA: CSLI Publications.
- Dalrymple, Mary, John J. Lowe & Louise Mycock. 2019. *The Oxford reference guide to Lexical Functional Grammar*. Oxford: Oxford University Press. DOI: 10.1093/oso/9780198733300.001.0001.
- Dalrymple, Mary & Louise Mycock. 2011. The prosody-semantics interface. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '11 conference*, 173–193. Stanford, CA: CSLI Publications.
- Dalrymple, Mary & Irina Nikolaeva. 2011. *Objects and information structure* (Cambridge Studies in Linguistics). Cambridge, UK: Cambridge University Press.
- Dalrymple, Mary, Agnieszka Patejuk & Mark-Matthias Zymla. 2020. XLE+Glue – A new tool for integrating semantic analysis in XLE. In Miriam Butt & Ida Toivonen (eds.), *Proceedings of the LFG '20 conference*, 89–108. Stanford, CA: CSLI Publications.
- Davis, Anthony R. 2001. *Linking by types in the hierarchical lexicon*. Stanford, CA: CSLI Publications.

- Davis, Anthony R. & Jean-Pierre Koenig. 2021. The nature and role of the lexicon in HPSG. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- De Kuthy, Kordula. 2021. Information structure. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Donohue, Cathryn & Ivan A. Sag. 1999. Domains in Warlpiri. Unpublished manuscript, Stanford University. <https://www.academia.edu/download/30754990/donohue-sag99.pdf>.
- Enderton, Herbert B. 1977. *Elements of set theory*. New York: Academic Press.
- Fang, Ji & Peter Sells. 2007. A formal analysis of the verb copy construction in Chinese. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '07 conference*, 198–213. Stanford, CA: CSLI Publications.
- Findlay, Jamie Y. 2016. Mapping theory without argument structure. *Journal of Language Modelling* 4(2). 293–338. DOI: 10.15398/jlm.v4i2.171.
- Francez, Nissim & Shuly Wintner. 2012. *Unification grammars*. Cambridge, UK: Cambridge University Press.
- Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum & Ivan A. Sag. 1985. *Generalized phrase structure grammar*. Cambridge, MA: Harvard University Press.
- Haug, Dag & Tatiana Nikitina. 2012. The many cases of non-finite subjects: The challenge of “dominant” participles. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 292–311. Stanford, CA: CSLI Publications.
- Hellan, Lars. 2019. Construction-based compositional grammar. *Journal of Logic, Language and Information* 28. 101–130. DOI: 10.1007/s10849-019-09284-5.
- Höhle, Tilman N. 1999. An architecture for phonology. In Robert D. Borsley & Adam Przepiórkowski (eds.), *Slavic in Head-Driven Phrase Structure Grammar*, 61–90. Stanford, CA: CSLI Publications. Reprinted in Müller, Reis & Richter 2019.
- Johnson, Mark. 1988. *Attribute-value logic and the theory of grammar*. Stanford, CA: CSLI Publications.
- Kaplan, Ronald M. 1987. Three seductions of computational psycholinguistics. In Peter Whitelock, Mary McGee Wood, Harold L. Somers, Rod Johnson & Paul Bennett (eds.), *Linguistic theory and computer applications*, 149–188. London: Academic Press. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 339–367).

Adam Przepiórkowski

- Kaplan, Ronald M. 1989. The formal architecture of Lexical-Functional Grammar. *Journal of Information Science and Engineering* 5. 305–322. Revised version published in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 7–27).
- Kaplan, Ronald M. 1995. The formal architecture of Lexical-Functional Grammar. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III & Annie Zaenen (eds.), *Formal issues in Lexical-Functional Grammar*, 7–27. Stanford, CA: CSLI Publications. Earlier version published as Kaplan (1989).
- Kaplan, Ronald M. 2017. Preserving grammatical functions in LFG. In Victoria Rosén & Koenraad De Smedt (eds.), *The very model of a modern linguist – In honor of Helge Dyvik*, vol. 8 (Bergen Language and Linguistics Studies), 127–142. Bergen: Bergen Language & Linguistics Studies (BeLLS). DOI: [10.15845/bells.v8i1.1342](https://doi.org/10.15845/bells.v8i1.1342).
- Kaplan, Ronald M. 2019. Formal aspects of underspecified features. In Cleo Condoravdi & Tracy Holloway King (eds.), *Tokens of meaning: Papers in honor of Lauri Karttunen*, 349–369. Stanford, CA: CSLI Publications.
- Kaplan, Ronald M. & Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 173–281. Cambridge, MA: The MIT Press. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 29–130).
- Kathol, Andreas. 1995. *Linearization-based German syntax*. Columbus, OH: Ohio State University. (Doctoral dissertation).
- Kathol, Andreas. 2000. *Linear syntax*. Oxford: Oxford University Press.
- Kathol, Andreas & Carl Pollard. 1995. Extraposition via complex domain formation. In *Proceedings of the 33rd annual meeting of the Association for Computational Linguistics*, 174–180. Cambridge, MA. DOI: [10.3115/981658.981682](https://doi.org/10.3115/981658.981682).
- Kepser, Stephan. 2004. On the complexity of RSRL. In *Electronic notes in theoretical computer science*, vol. 53, 146–162. Amsterdam: Elsevier. DOI: [10.1016/s1571-0661\(05\)82580-0](https://doi.org/10.1016/s1571-0661(05)82580-0).
- King, Paul John. 1989. *A logical formalism for Head-driven Phrase Structure Grammar*. Manchester: University of Manchester. (Doctoral dissertation).
- King, Paul John. 1999. Towards truth in HPSG. In Valia Kordoni (ed.), *Tübingen studies in Head-Driven Phrase Structure Grammar* (Arbeitspapiere des Sonderforschungsbereichs 340, Bericht Nr. 132), 301–352. Tübingen: Universität Tübingen.
- Kroeger, Paul R. 1993. *Phrase structure and grammatical relations in Tagalog*. Stanford, CA: CSLI Publications.
- Kubota, Yusuke. 2021. HPSG and Categorical Grammar. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase*

- Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Kuhn, Jonas. 2001. Resource sensitivity in the syntax-semantics interface: Evidence from the German split NP construction. In W. Detmar Meurers & Tibor Kiss (eds.), *Constraint-based approaches to Germanic syntax* (Studies in Constraint-Based Lexicalism), 177–216. Stanford, CA: CSLI Publications.
- Lambek, Joachim. 1958. The mathematics of sentence structure. *The American Mathematical Monthly* 65(3). 154–170. DOI: [10.2307/2310058](https://doi.org/10.2307/2310058).
- Lowe, John J. 2016. Clitics: Separating syntax and prosody. *Journal of Linguistics* 52. 375–419.
- Lowe, John J. & Joseph Lovestrand. 2020. Minimal phrase structure: A new formalized theory of phrase structure. *Journal of Language Modelling* 8(1). 1–51. DOI: [10.15398/jlm.v8i1.247](https://doi.org/10.15398/jlm.v8i1.247).
- Malouf, Robert. 1998. *Mixed categories in the hierarchical lexicon*. Stanford, CA: Stanford University. (Doctoral dissertation).
- McCawley, James D. 1968. Concerning the base component of a transformational grammar. *Foundations of Language* 4(3). 243–269.
- Müller, Stefan. 2021. Constituent order. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Müller, Stefan, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.). 2021. *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Müller, Stefan & António Branco. 2021. Anaphoric binding. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Müller, Stefan, Marga Reis & Frank Richter (eds.). 2019. *Beiträge zur deutschen Grammatik: Gesammelte Schriften von Tilman N. Höhle*. Berlin: Language Science Press.
- Nykiel, Joanna & Jong-Bok Kim. 2021. Ellipsis. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Patejuk, Agnieszka & Adam Przepiórkowski. 2016. Reducing grammatical functions in LFG. In Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King & Stefan Müller (eds.), *Proceedings of the joint 2016 conference on Head-Driven Phrase Structure Grammar and Lexical Functional Grammar*, 541–559. Stanford, CA: CSLI Publications.

Adam Przepiórkowski

- Pereira, Fernando C. N. & David H. D. Warren. 1983. Parsing as deduction. In *Proceedings of the 21st annual meeting of the Association for Computational Linguistics*, 137–144. DOI: 10.3115/981311.981338.
- Pollard, Carl. 1997. The nature of constraint-based grammar. *Linguistic Research* 15. 1–18. <http://isli.khu.ac.kr/journal/content/data/15/1.pdf>.
- Pollard, Carl. 1999. Strong generative capacity in HPSG. In Gert Webelhuth, Jean-Pierre Koenig & Andreas Kathol (eds.), *Lexical and constructional aspects of linguistic explanation*, 281–297. Stanford, CA: CSLI Publications.
- Pollard, Carl. 2001. Cleaning the HPSG garage: Some problems and some proposals. Unpublished manuscript, Ohio State University. [http://utkl.ff.cuni.cz/~rosen/public/cp\\_garage.ps](http://utkl.ff.cuni.cz/~rosen/public/cp_garage.ps).
- Pollard, Carl. 2014. Type-logical HPSG. In Gerhard Jäger, Paola Monachesi, Gerald Penn & Shuly Wintner (eds.), *Proceedings of Formal Grammar 2004*, 111–128. Stanford, CA: CSLI Publications.
- Pollard, Carl & Ivan A. Sag. 1987. *Information-based syntax and semantics*. Stanford, CA: CSLI Publications.
- Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press & CSLI Publications.
- Przepiórkowski, Adam. 2016. How *not* to distinguish arguments from adjuncts in LFG. In Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King & Stefan Müller (eds.), *Proceedings of the joint 2016 conference on Head-Driven Phrase Structure Grammar and Lexical Functional Grammar*, 560–580. Stanford, CA: CSLI Publications.
- Przepiórkowski, Adam & Agnieszka Patejuk. 2012. The puzzle of case agreement between numeral phrases and predicative adjectives in Polish. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 490–502. Stanford, CA: CSLI Publications.
- Pullum, Geoffrey K. 2007. The evolution of model-theoretic frameworks in linguistics. In James Rogers & Stephan Kepser (eds.), *Model-theoretic syntax at 10*, 1–10.
- Pullum, Geoffrey K. 2019. What grammars are, or ought to be. In Stefan Müller & Petya Osenova (eds.), *Proceedings of the 26th international conference on Head-Driven Phrase Structure Grammar*, 58–79. Stanford, CA: CSLI Publications.
- Pullum, Geoffrey K. & Barbara C. Scholz. 2001. On the distinction between model-theoretic and generative-enumerative syntactic frameworks. In Philippe de Groote, Glyn Morrill & Christian Retoré (eds.), *Logical aspects of computational linguistics: 4th international conference*, vol. 2099 (Lecture Notes in Artificial Intelligence), 17–43. Berlin: Springer. DOI: 10.1007/3-540-48199-0\_2.



- Reape, Mike. 1992. *A formal theory of word order: A case study in West Germanic*. Edinburgh: University of Edinburgh. (Doctoral dissertation).
- Reape, Mike. 1996. Getting things in order. In Harry Bunt & Arthur van Horck (eds.), *Discontinuous constituency*, 209–253. Berlin: Mouton.
- Richter, Frank. 1999. RSRL for HPSG. In Valia Kordoni (ed.), *Tübingen studies in Head-Driven Phrase Structure Grammar* (Arbeitspapiere des Sonderforschungsbereichs 340, Bericht Nr. 132), 74–115. Tübingen: Universität Tübingen.
- Richter, Frank. 2004. *A mathematical formalism for linguistic theories with an application in Head-Driven Phrase Structure Grammar*. Tübingen: Universität Tübingen. (Doctoral dissertation).
- Richter, Frank. 2007. Closer to the truth: A new model theory for HPSG. In James Rogers & Stephan Kepser (eds.), *Model-theoretic syntax at 10*, 99–108.
- Richter, Frank. 2021. Formal background. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Richter, Frank & Manfred Sailer. 1997. Underspecified semantics in HPSG. In Harry Bunt & Reinhard Muskens (eds.), *Computing meaning* (Studies in Linguistics and Philosophy), 95–112. Dordrecht: Springer. DOI: 10.1007/978-94-011-4231-1\_5.
- Richter, Frank & Manfred Sailer. 1999. LF conditions on expressions of Ty2: An HPSG analysis of negative concord in Polish. In Robert D. Borsley & Adam Przepiórkowski (eds.), *Slavic in Head-Driven Phrase Structure Grammar*, 247–282. Stanford, CA: CSLI Publications.
- Rogers, James. 1997. “Grammarless” phrase structure grammar. *Linguistics and Philosophy* 20. 721–746.
- Rogers, James. 1998. *A descriptive approach to language-theoretic complexity*. Stanford, CA: CSLI Publications.
- Sag, Ivan A. 2012. Sign-Based Construction Grammar: An informal synopsis. In Hans C. Boas & Ivan A. Sag (eds.), *Sign-Based Construction Grammar*, 61–197. Stanford, CA: CSLI Publications.
- Sailer, Manfred. 2003. *Combinatorial semantics and idiomatic expressions in Head-driven Phrase Structure Grammar*. Tübingen: Universität Tübingen. (Doctoral dissertation).
- Simpson, Jane. 1991. *Warlpiri morpho-syntax: A lexicalist approach*. Dordrecht: Kluwer Academic Publishers.
- Snijders, Liselotte. 2015. *The nature of configurationality in LFG*. Oxford: University of Oxford. (D.Phil. Thesis).
- Søgaard, Anders & Martin Lange. 2009. Polyadic dynamic logics for HPSG parsing. *Journal of Logic, Language and Information* 18. 159–198.

*Adam Przepiórkowski*

- Wechsler, Stephen & Ash Asudeh. 2021. HPSG and Lexical Functional Grammar. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*. Forthcoming. Berlin: Language Science Press.
- Wechsler, Stephen & Larisa Zlatić. 2003. *The many faces of agreement*. Stanford, CA: CSLI Publications.
- Wedekind, Jürgen & Ronald M. Kaplan. 2020. Tractable Lexical-Functional Grammar. *Computational Linguistics* 46(2). 515–569. DOI: [10.1162/coli\\_a\\_00384](https://doi.org/10.1162/coli_a_00384).
- Wescoat, Michael T. 2002. *On lexical sharing*. Stanford, CA: Stanford University. (Doctoral dissertation).
- Zaenen, Annie & Ronald M. Kaplan. 1995. Formal devices for linguistic generalizations: West Germanic word order in LFG. In Jennifer S. Cole, Georgia M. Green & Jerry L. Morgan (eds.), *Linguistics and computation*, 3–27. Stanford, CA: CSLI Publications. Reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 215–240).
- Zweigenbaum, Pierre. 1988. *Attributive adjectives, adjuncts and cyclic f-structures in Lexical-Functional Grammar*. DIAM Rapport Interne RI-58a. Paris: Département Intelligence Artificielle et Medecine.





# Handbook of Lexical Functional Grammar

Set blurb on back with \BackBody{my blurb}

