Handbook of Lexical Functional Grammar

Edited by

Mary Dalrymple

Empirically Oriented Theoretical Morphology and Syntax

Empirically Oriented Theoretical Morphology and Syntax

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. Bowern, 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.

ISSN: 2366-3529

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

ISBN: no digital ISBN
no print ISBNs!
ISSN: 2366-3529
no DOI
Source code available from www.github.com/langsci/312
Collaborative reading: 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: XeLaTeX

Language Science Press
xHain
Grünberger Str. 16
10243 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

Contents

Chapter 1

Information structure

Annie Zaenen
Stanford University

The first section of this chapter gives an overview of the conceptual discussions in information structure (IS), and the second section describes the LFG work in the area. The third section intends to be an exhaustive overview of the LFG work on IS.

In LFG, attention to information structure has led to many interesting language specific studies but, contrary to the situation in, for instance, syntax, there is no generally accepted view of either the distinctions needed or the terminology to be adopted. Given this situation, we start with a general, not LFG-specific, overview of conceptual discussions in IS (Section 1). The hope is that this will alert the reader to check which notion of say, topic, focus or contrast, is used in the LFG contribution they happen to be reading. In the second section (Section 2), we describe the general lines of the LFG work, highlighting some of the concepts that are often appealed to and the major proposals that have been made about how IS should be integrated in the LFG architecture. The third section (Section 3) gives exhaustive thumbnails of the LFG work on IS. The overview in this chapter does not include historical studies. They will be covered in Booth & Butt forthcoming [this volume] (LFG and historical linguistics).

1 What is Information Structure?

Information Structure looks at how a speaker-writer presents linguistically encoded information to the hearer-reader. It studies the sentence internal aspects of this organization, while Discourse Structure (DS) studies the overall organization of bigger units of a text. The term INFORMATION STRUCTURE (IS) was coined by Halliday (1967); Chafe (1976) introduced the term INFORMATION PACKAGING,

Annie Zaenen. Forthcoming. Information structure. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

which is also used by Vallduví (1992). Recently it has been generally accepted that it is desirable to keep DS and IS separate, but it is often not possible to describe IS without making some assumptions about DS: for instance, the notion of sentence topic and discourse topic are related and often not kept clearly separated. Both DS and IS are generally considered to be part of pragmatics and different from semantics, which is mainly concerned with INFORMATION CONTENT while DS and IS are concerned with INFORMATION MANAGEMENT (cf. Krifka 2007). But, again, certain aspects of information content and information management are closely intertwined, e.g. scope phenomena, the notion of predication, and pronoun interpretation (see e.g. Reinhart 1981, also King & Zaenen 2004).

IS distinctions can be realized through *prosody*, *word order* and/or *morphology*. In this section we mainly discuss word order and morphology; for prosody, see Bögel forthcoming [this volume].

IS entities can be talked about as being linguistic in nature, e.g. NPs, VPs, etc. or they can be thought of as psychological. In most cases the only thing we can access and study are the linguistic reflexes of psychological states. This leads to terminological confusion. E.g. is a topic a textual entity in the sentence or is it the entity/state of affairs the speaker in this utterance intends to talk about, the denotation? Reinhart's discussion, for instance, is very much in terms of linguistic entities: all the NPs in a sentence are possible topics, and the new information will be added to the file card for the NP or a new file card will be created.

In the discussion of topic, the ambiguity is often not too harmful. It leads to more confusion in the discussion of focus. Consider the sentence *John washed it*, as an answer to *What happened to the car?* Here *John washed* is taken to be the focus but in most syntactic theories this is not a constituent at any level, the exception being Steedman's Categorial Grammar framework (see e.g. Steedman 2000).

Lambrecht (1994) and Vallduví (1992) state explicitly that they see topic and focus as psychological and that the linguistic entities should be termed TOPIC and FOCUS EXPRESSIONS. They appeal to a pragmatic notion of common ground[1] in which speaker and hearer's intentions play a role. This relates to the issue of how IS fits into a full-fledged formal representation of text. In general that is left

[1]The original notion comes from Stalnaker (1970), who defines it as the set of pragmatic presuppositions shared by speaker and hearer at the moment of the utterance of the sentence. For him a proposition is pragmatically presupposed when the participants in the discourse 'take its truth for granted', and 'assume that others involved in the context do the same'. It is not totally clear that Lambrecht has the same idea. He states "'To have knowledge of a proposition' is understood here in the sense of 'to have a mental picture of its denotatum', not in the sense of 'to know its truth'" (Lambrecht 1994: 44; see Dryer 1995 for an extensive discussion.)

rather vague. The most common metaphor is that the IS gives instructions to update file cards à la Heim (1982) or Kamp (1981). For Lambrecht (1994), the pragmatic presupposition and assertion are sets of propositions, and topic and focus structure are defined in relation to them. How all this is ultimately represented in the human mind, is definitely outside of this overview. For convenience's sake, I will assume that speakers and hearers have a structured representation of the discourse they are engaged in, and I will refer to this as the DENOTATION of the linguistic entities that encode IS.

IS, viewed as part of pragmatics, assumes that information transmission relies on the speaker and the hearer having a common ground from which the interaction proceeds, with the speaker-writer introducing information not yet known by the hearer-reader. The interlocutors are in a particular information state that they intend to change through the interaction. An utterance, then can be divided into two parts, one that links it to what precedes and one that introduces new information. Under this bipartite subdivision, the TOPIC, THEME or BACKGROUND is the part that relates it to the preceding discourse and the COMMENT, RHEME or FOCUS advances the discourse by providing new information. Even ignoring the notion of DISCOURSE TOPIC and the existence of ALL FOCUS utterances, both sets of notions are unclear and defined differently by different authors. A useful way to get a grip on the main distinctions is to start from Dahl's (1974) observation (see also Jacobs 1984) that, assuming a context in which what John drinks is at issue, a sentence such as *John drinks beer* can simultaneously be analyzed as having *John* as its topic and *drinks beer* as the comment or *John drinks* as the background and *beer* as the (narrow) focus. The sentence is about *John* (the topic), so the rest is comment, but the new information, the focus, is *beer*, as we already know that John drank something. Researchers who are interested in focus often see the main division as BACKGROUND-FOCUS, whereas those interested in topic see it more often as TOPIC-COMMENT. The two views are combined in the tripartite proposal of Vallduví, who proposes LINK-FOCUS-TAIL, where LINK and TAIL together form the GROUND.

Apart from these discrete views, there are views that see information structure as gradient: for instance, the Prague school (e.g. Firbas 1975) describes topics as the material lowest in "communicative dynamism" where the latter is determined by three parameters: linear order, semantic considerations (e.g. the type of the verb), and the degree of context dependency. These views are not often referred to in the LFG literature but some researchers have proposed hierarchical analyses with respect to the notions of activation/salience (see Andréasson 2009, O'Connor 2006) or topic-worthiness (Dalrymple & Nikolaeva 2011)

Other authors have proposed to further analyze notions such as topic and focus as feature bundles. This allows for more fine-grained distinction. We will discuss the ones proposed in LFG further in Section 2.1.

There is no accepted view of IS in LFG, but the views that have had the most influence are Lambrecht (1994),Vallduví (Vallduví 1992, Vallduví & Vilkuna 1998, Vallduví & Engdahl 1996), and Reinhart (1981). Only a few LFG proposals address the conceptual issues whole-sale.

In what follows I will review some IS notions in more detail and then discuss the LFG work in the area.

1.1 Notions of topic

One of the most influential proposals is that of ABOUTNESS TOPIC, so named by Reinhart (1981), but the idea goes back at least as far as Kuno (1987) (see also Dahl 1974). The link to the previous discourse that is assumed here is rather broad: the referent of a topic is presupposed to exist (based on Strawson 1964). Reinhart, following Kuno, uses locutions such as *X says about Y*, *Speaking of X*, *As for X* as tests to distinguish the topic from other NPs in the sentence. It is recognized that these tests do not work very well. *Speaking of X*, for instance, is an expression notoriously used to change the topic, so it cannot be used as a diagnostic for continuing topics.

Reinhart uses the file card metaphor (cf. Heim 1982) to explain what a topic does: indicate where the hearer should store the information contained in the sentence. The proposal does not distinguish between continuing topics and switch topics (see below) and makes the explicit assumption that every sentence has only one topic. For instance, for the sentence *All crows are black*, the information provided is classified under crows and understood as an assertion about the set of all crows. (So, our natural way to assess it will be to check the members of this set and see if any of them are not black, rather than checking the non-black things we know to see if any of them are a crow.) This view seems to be based on an overly literal conception of a file card. With a discourse fragment like *What about John and Mary? They got married but he doesn't love her* (adapted from Lambrecht 1994), it seems difficult to claim that this information is stored only under John rather than (also) under Mary.

The one topic idea is explicitly rejected by Lambrecht (1994) who sees the difference that Reinhart (1981) sees between topics and non-topic definites as a difference in salience. This view has led to the proposal to distinguish between primary and secondary topics (see below). Lambrecht also rejects Reinhart's semantic notion of presupposition. For Reinhart, following Strawson (1964), topics

are presupposed to exist. Lambrecht's notion of presupposition is pragmatic and does not require this. For him, the most important pragmatic articulation of a structured meaning is that between PRAGMATIC PRESUPPOSITION and PRAGMATIC ASSERTION (comment), where the pragmatic presupposition is assumed by the speaker to be equally assumed by the hearer and the pragmatic assertion is what the hearer is expected to assume after having heard the sentence. For Lambrecht *nobody* can be a topic. Most authors, however, confine the notion of topic to referents that can be expressed by definites or specific indefinites. But not everybody agrees with this, see e.g. Endriss (2009) and Gécseg & Kiefer (2009).

Vallduví's notion of LINK is close to that of topic, but it is explicitly restricted to elements in first position. He sees this first position[2] as a requirement for them to act as address pointer, instructing the hearer to go to an address or card in the sense of Heim (1982). When the elements in the sentence do not require a pointer to a new address, they are not LINKS (topics) but TAILS. Tails correspond more or less to what others have called SECONDARY or CONTINUING TOPICS. Vallduví's notion of link is restricted to what have been called SWITCH TOPICS: topics that are different from the topic of the previous sentence.[3]

As this short discussion shows, researchers have felt it to be useful to distinguish topics depending on the function of the element they are in an anaphoric relation with in the previous discourse. (The discussion tends to be restricted to the previous sentence.) The antecedent can be a topic, in which case we have a CONTINUING TOPIC or it can be a focus, in that case we have a SWITCH TOPIC (corresponding to Vallduví's link). It has been claimed that some constructions explicitly signal switch topics.[4] With the notion of CONTINUING TOPIC we come close to that of DISCOURSE TOPIC. The two might be distinguished in the sense that a continuing topic has to be the topic of the immediately preceding discourse unit (e.g. sentence), whereas the discourse (or FAMILIAR) topic can be broader, but not all authors make that distinction. Some authors will reserve the term ABOUTNESS TOPIC for SWITCH TOPIC and hence distinguish them from CONTINUING TOPICS (see e.g. Frascarelli & Hinterhölzl 2007, who show that in Italian and German the types have a different linguistic realization). CONTINUING topics are often

[2]He allows for two topics under the condition that they precede all other elements in the sentence.

[3]The discussion in section 5.1.1 of Vallduví & Engdahl (1996) shows that the alignment between switch topic and first position cannot be right for all languages. It also shows that restricting the notion of topic to switch topic is awkward, as Swedish examples show this is also a position of continuing topics, a notion that is introduced as an addition to the Vallduví classification in the paper.

[4]e.g. first position elements in Catalan as analyzed in Vallduví (1992)

not overtly expressed (e.g. in so-called pro-drop languages). Some authors will consider these sentences as topicless, others will assume the understood topic to be part of the IS of the sentence.

A further distinction is invoked with the term CONTRASTIVE topic. Contrast is most often invoked with respect to foci and we will discuss the concept further in Section 1.2.1. Here we point out that topic expressions can contain elements that are new as well as old ones. E.g. *What about cars? Which ones do you like? Fast cars I like.* The *cars* are topical but *fast* is new information. Contrastive topics are seen as implying that a set of alternatives is active in the speaker mind. They can be linked to the notion of D-LINKING (Pesetsky 2007). D-linked elements are defined (following Büring 2003) as related to the question under discussion (QUD). This, again, brings discourse topics into the discussion of sentence topics.

The types of topics mentioned above are typically expressed as NPs but there is a kind of topic that is most often realized as a PP. It occurs in so-called "all focus" sentences such as *In California, there are often forest fires.* Here, the initial element restricts the range of the rest of the sentence. These elements have been dubbed STAGE TOPICS by Erteschik-Shir (2007) but they have been discussed earlier, e.g. Gundel (1974). Erteschik claims there is a silent stage topic in all-focus sentences that don't have an overt topic. Some researchers seem to think the stage topic only occurs in sentences that have no other topic, but for others, stage topics can co-occur with ABOUTNESS topics. Bentley & Cruschina (2018) discuss the specific lexical-semantic restrictions on all-focus constructions in Romance languages. LFG researchers have not worked on stage topics but Szűcs (2017) discusses English adverbials that could be considered as, in part, falling in that class.

1.1.1 Accessibility Hierarchies

Apart from the aboutness tests proposed by Reinhart, some proposals in IS appeal to notions, such as salience or topic-worthiness, that distinguish among the entities that are assumed to be in the discourse participants' consciousness at the moment a new utterance gets produced or heard. Following Chafe (2007), who introduced the idea, various hierarchies or, at least, classifications have been proposed (see e.g. Givón 1983, Ariel 1988, Lambrecht 1994, Erteschik-Shir 2007). Within LFG the ones that are referred to are Prince's (1981) notions of DISCOURSE OLD/NEW and HEARER OLD/NEW, EVOKED, INFERRABLE (see also Ward & Birner 2001) and Gundel's GIVENNESS HIERARCHY (e.g. Gundel et al. 1993).

Prince's categories are about entities. If they have been mentioned in the discourse, they are discourse old, if not they are discourse new (or inferable). Refer-

ents may also be old/new with respect to (the speaker's beliefs about) the hearer's beliefs. In feature decomposition approaches in LFG, following Choi (1996), the +NEW feature corresponds to Prince's notion of discourse new. Lambrecht (1994) proposes a connection between Prince's notions and topics (his Topic Accessibility Scale) but notes that inherent semantic factors such as animacy may also play a role.

The GIVENNESS HIERARCHY of Gundel et al. proposes six ordered cognitive statuses (in focus > activated > familiar > uniquely identifiable > referential > type identifiable). They determine the form of referring expressions and are assumed to correspond to the status of the referent in the memory of the discourse participants. For instance, for English:

- TYPE IDENTIFIABLE: necessary for the appropriate use of any nominal expression
- REFERENTIAL: speaker refers to a particular object, hearer needs to know the referent, necessary for appropriate use of definite expressions
- UNIQUELY IDENTIFIABLE: necessary for the appropriate use of *the*
- FAMILIAR: necessary for the uses of personal and definite demonstrative pronouns
- ACTIVATED: necessary for all pronominal forms and sufficient for demonstrative *that* and stressed personal pronouns
- IN FOCUS: necessary for the use of zero and unstressed pronouns.

Especially the two last items on the scale have been used to argue for differences in status between elements that are not treated equally: cognitive accessibility relates to how prominent the entity is memory. This does not in itself determine whether it will be a topic or a focus. The accessible elements are all topic-worthy but depending on their place on the scale, they will require a different linguistic expression.

1.1.2 Other Hierarchies

The topic-worthiness of a discourse element has also been claimed to be influenced by the prominence features that play a role in the REFERENTIAL HIERARCHY (see Silverstein 1976, also Dixon 1994), such as person, definiteness, animacy. This hierarchy has played a more important role in the studies about the alignment of Grammatical Functions but, as Simpson (2012) following Bickel (2012) observes,

a hierarchy with pronouns on one end and full NPs on the other is bound to have something to do with IS.

In recent literature, it is generally accepted that topic-worthiness or accessibility are not enough to guarantee topichood and that ABOUTNESS is the crucial factor. It is, however, not clear how aboutness can be detected. The decision to construct a sentence/utterance about a particular sentence topic seems to be a decision that the speaker/writer makes which is constrained, but not uniquely determined, by the previous discourse. In some languages this choice needs always to be clearly marked, while in others that is not the case and a specific marking may be absent or optional.

1.2 Focus and related notions

The focus is (or is part of) what is informationally new in a sentence, what is not assumed to be common ground between the hearer and the speaker at the moment that the sentence is uttered. A common proposal is that the focus can be found by considering what question the sentence could be an answer to. The focus is what replaces the wh-term in the question.[5] A typical set of question-answer pairs is the following:

(1)
- Q: What did Mary do?
- A: She [washed the car].

(2)
- Q: What did Mary wash?
- A: She washed [the car].

(3)
- Q. Who washed the car?
- A: [Mary] washed the car.

(4)
- Q: What happened to the car?
- A: [Mary washed] it.

(5)
- Q: What happened?
- A: [Mary washed the car].

(6)
- Q: What did Mary do with the car?

[5]This test runs into problems when a particular language has a special marking for answers to questions that distinguishes those from other arguable foci. Another more general problem with the test is that full answers to wh-questions are often unnatural. Most of the topical information would be silent. These versions are, however, rather uninformative when the test is used to probe word order constraints.

- A: She [washed] it.

The focus of each answer is the material in square brackets. Lambrecht (1994) distinguished three notions of focus structure: PREDICATE FOCUS (1 and, presumably, 4 and 6), ARGUMENT FOCUS (2 and 3) and SENTENCE FOCUS (5)

The question-answer approach tends to tie the IS notion of focus to alternative sets as used in semantics, as the meaning of wh-questions can be considered as the set of possible answers (Hamblin 1973), or one can think about the alternative sets of Rooth's (1992) view on focus and focus particles. The problem with appealing to these semantic notions is that one can get bogged down by the issue of how explicit these alternatives have to be. It is clear that, in a certain sense, every assertion is made against the background of all other possible assertions that could have been made at that particular moment in the discourse, but that doesn't mean that one can list/define a set of alternatives (see below).

It is often claimed that different types of foci are distinguished by the degree to which the set of alternatives has been made explicit. In the example above, the syntax is unremarkable and the stress pattern is what one would find in normal narrative text. The foci in this example are called INFORMATION or COMPLETIVE or IDENTIFICATION FOCI. They are often expressed in what is thought of as 'neutral' or 'default' syntax. In some languages, specific constructions allow the speaker to signal whether she has a particular set of alternatives in mind, or even whether she wants to convey that only one option is possible. This has led to the distinction between the foci above and CONTRASTIVE and/or EXHAUSTIVE FOCI. Moreover, EXCLUSIVE foci can be distinguished from EXHAUSTIVE foci because exclusive foci should exclude some alternatives while EXHAUSTIVE foci exclude all alternatives. Some researchers have proposed many more subdivisions of focus. Dik et al. (2008) distinguishes COMPLETIVE focus, which is NON-CONTRASTIVE, from all other CONTRASTIVE forms: PARALLEL, SELECTIVE and three types of corrective foci: EXPANDING, RESTRICTING, and REPLACING. Most of these subtypes can be seen as specifying relations between the set of alternatives and the focus.[6]

[6]Not all researchers distinguish clearly between the FOCUS DOMAIN and FOCUS EXPONENT. For instance, in the English examples above, focus is normally indicated with pitch accent. As discussed at least since Jackendoff (1972), the nuclear stress rule will assign stress to the final constituent of a focus, while the focus itself can be projected up any higher constituent, so 1, 2 and 5 will get the same stress assignment but have different foci. Even when phenomena such as FOCUS PROJECTION are recognized, researchers tend to concentrate their attention on FOCUS MARKERS and are often not very clear on what constitutes the FOCUS DOMAIN. For some it is, actually, the markers that deserve the term focus. In this overview I am assuming that the conceptual category can be distinguished from its realizations.

In running text, it is not always clear what the question is. Consider the second sentence of the example (adapted from Vallduví & Engdahl 1996) below.

(7)
- Mary bought a book yesterday morning.
- She read it in the afternoon.

The second sentence can be seen as an answer to the question *What did she do then?* as well as an answer to the question *What did she do with the book?* In the first case, the (denotation of) the VP is focal. In the second case, both *she* and *it* are topics and the rest of the material is focal. Written material, especially reduced by a window of at most two sentences, is prone to being pragmatically ambiguous. By turning the text into a set of questions and answers one imposes an interpretation that reduces the ambiguity. Looking at larger pieces of text might also help in figuring out what the right question is in a particular context but that leads to discourse analysis as distinct from information structure.

1.2.1 Contrast

Although there are many subdivisions of focus, the main dividing line seems to be between non-contrastive and contrastive focus. Once this line has been drawn, however, one realizes that contrast is not only relevant for focus but also for topic. Take, for instance, an exchange like the following:

(8)
- Q. Which foreign languages do your children speak?
- A. Anna speaks English and Maria speaks German.

In languages such as English there will be CONTRASTIVE STRESS on *Anna* and *Maria* as well as on *English* and *German*. Some researchers see this as a reason to adopt contrast as an independent notion and propose to distinguish between ±CONTRASTIVE topics and foci.

Here, again, a confusing factor is that contrast can be used to refer to an abstract category or to a linguistic signal, e.g. a particular stress pattern. Contrastive focus, then, can mean that the focus has some special stress or pitch pattern (dubbed Kontrast by Vallduví & Vilkuna 1998) or it can mean that the focus element contrasts with other elements that could fill the same position in the sentences by being an alternative to these elements. When seen as the latter, the notion is as confusing as that of focus and, in fact, it is difficult to see a difference between the two.

The various notions of contrast are discussed in Repp (2016). She distinguishes:

(9) a. Restricted, contextually clearly identifiable set of alternatives: *John, Pete and Josie all offered help. I asked John.* John would be marked for contrast (É. Kiss 1998).

b. Alternatives must be in the sentence.

c. Substitution of alternative must create false statement (e.g. Neeleman & Vermeulen 2013).

d. Alternatives always contrast, simply by being different (alternative semantics) (Vallduví & Vilkuna 1998, also Krifka 2007)

e. Interlocutors' belief systems: unexpected, remarkable (Frey 2006).

These are, by and large, the notions of set of alternatives used in discussions of focus and, at the limit, we find a notion *unexpected, remarkable* that seems to depend on the speaker's frame of mind without being independently detectable. comma after "notion" Moreover, as we have seen above, contrasts are not limited to foci: they can also occur with topics. Some researchers have bit the bullet and consider contrast to be the correct notion, e.g. Kruijff-Korbayová & Steedman (2003). Under that view, contrast within the topic indicates that part of the topic is actually focal (Krifka 2007; Erteschik-Shir 2007). Others, however, see contrast as an additional distinction, and much work in LFG takes that approach. It is a view that has been argued for explicitly by e.g. Neeleman & Vermeulen (2013) who distinguish between contrastive and noncontrastive topics as well as contrastive and noncontrastive foci. They illustrate their approach with word order data in Dutch but claim that in other languages it can be detected through prosodic marking. As said above, their notion of contrast relies on the judgments that in contrastive contexts, the substitution of alternatives leads to false statements.

1.2.2 Relational newness

For researchers who propose to see CONTRAST as a feature that can belong to both topics and foci, the question remains: what is the characteristic that distinguishes focus from topic? Choi (1996), among others, proposes that foci have to be DISCOURSE NEW in the sense of Prince (1981). This view is, however, contested by e.g. Lambrecht (1994) and Gundel (1974), who draws attention to examples such as (adapted from Lambrecht 1994):

(10)
- Last night Anne and Paul were bored.
- They hesitated between going to the neighborhood restaurant or going to the new movie at the Rex.
- Finally they went to the movie.

(The denotation of) *last night* in the first sentence can be considered to be a stage topic and the rest of the material in that sentence is focal. In the second and the third sentence, the aboutness topic is (the denotation of) *they* (=Anne and Paul). The second and the third sentence (as well as the first one) establish new relations between these referents and the rest of the sentence. In the third sentence, *the movie* and even *going to the movie* are as much old information as *they*. What is new is the relation between the elements. The referents of the NPs have become part of the common ground between speaker and hearer before the sentence is uttered. They are not DISCOURSE NEW in Prince's sense or REFERENTIALLY NEW in Gundel's terminology. But, they are RELATIONALLY NEW in Gundel's terminology. Once it is determined that the topic is (the denotation of) *they*, the choice between the restaurant and the movie is the relevant relationally new information. One can thus see (the denotation of) *the movie* as the element about which a new relation is asserted and mark it as +NEW. It is the alternative that is chosen in opposition to all other possible choices.[7]

Note that under this view of newness, it is not immediately clear that (the denotation of) *they* is the topic: the relation between (the denotation of) *they* and (the denotation of) the rest of the sentence is also new. To determine the focus, one has to know already what the topic is, one has to know what we are adding information about. Here we are assuming that the first sentence leads to the question *What were they going to do about being bored?* and the second one to *Where did they go?*

A sentence can establish new relations between several different entities. In the second sentence below, the *give* relation is between three participants.

(11)
- Mary was wondering what she would give little Hansi: the candy bar or the chocolate chip cookie.
- She gave him the chocolate chip cookie.

Assuming that Mary is the topic, one can analyze the sentence as the answer to *What did Mary do?* in which case the denotation of the whole VP is the focus, or as the answer to *What did she give to him?*[8] in which case the denotation of the *chocolate chip cookie* is the focus. The second question presupposes that

[7]Some authors, most clearly, Lambrecht (1994), see the relation itself as new, and hence as the focus. Still others, e.g. Erteschik-Shir (2007), see the focus as a complex structure that can contain topical material.

[8]There are other possibilities which we ignore here. Our point is not an exhaustive analysis of this stretch of text.

both speaker and hearer already assume that Mary has given Hansi something, so the relational new information is that it is a chocolate chip cookie. What is the status of *him* under that analysis? Here it seems useful to remember the terminological ambiguity that was remarked upon in the introduction: is focus seen as a synonym of comment or is it to be opposed to background? When one assumes an opposition to background, it is reasonable to consider *him* to be part of the background and, possibly, as a special part of the background: a secondary topic.

The comment, too, can consist of more than just the focus. Consider the following question-answer pair:

(12)
- Q: Where was Mary?
- A: She was cooking potatoes in the kitchen

Here *in the kitchen* is the answer to the question and presumably the focus but *cooking potatoes* is also new information. Presumably it is part of the COMMENT. It can be seen as supplementary or COMPLETIVE information.[9]

2 The LFG approach

For LFG, there are two main issues related to IS: (1) what are the relevant distinctions to be made and how are these encoded and (2) how does IS interact with the LFG architecture. We discuss these in turn.

2.1 Feature decomposition

In the previous section we have seen that the notions of topic and focus, although generally accepted, are felt to be insufficient to encode all relevant IS distinctions. With respect to topic there seems to be a need to further distinguish between different levels of salience and/or topic-worthiness of the entities that are accessible to the discourse participants. With respect to focus there seems to be a

[9]What has been analyzed in the literature as focus marking is very heterogeneous. This has led some researchers, most prominently Matič & Wedgwood (2013), to question the assumption that the linguistic processes that are described in the literature as marking foci indeed have a uniform function. Hedberg (2006) already argues that several proposals about the relation between pitch accent and IS in English are not mutually compatible, nor are the proposals about *wa* in Japanese or those about *nun* in Korean. Matič & Wedgwood (2013) go further and try to show that many proposals invoking focus marking in various languages actually isolate markers whose functions are quite different and whose effect on the focus is only a byproduct of these functions.

need to distinguish between explicitly contrastive and not explicitly contrastive elements. Moreover, it has been observed that some topical information assumes the existence of subsets among which a choice has to be made. This has led many researchers and, especially many LFG researchers to use features, most often binary features, to describe the IS behavior of linguistic entities and to define notions such as topic and focus. There is at this point no closed set of such features. Descriptive studies in this vein would provide a good basis for more theoretical investigations if the terminology was constant and explicit. Unfortunately, this is not the case. The same labels are used for different concepts and often no clear definition is given that allows the reader to figure out which meaning of an ambiguous term is intended. There are, however, some general tendencies: many researchers follow Choi (1996) and Butt & King (1996) and decompose the notions of topic and focus with two binary features.

In the light of the preceding discussion, one might expect that the notion of contrast would be represented in these dichotomies. This is, however, not the case in the Choi (1996) and Butt & King (1996) proposals.

Choi proposes the features ±NEW and ±PROM, for 'prominence', as shown in Table 1. Her notion of NEW is Prince's (1981) discourse new. Above, we pointed out that that notion is based on what is mentioned in the discourse and, hence, is problematic for certain analyses of focus. Her notion of PROM collapses the distinction between contrastive and completive focus and that between tail and link in Vallduví's sense. She does not discuss explicitly what such a collapsed notion would correspond to intuitively.

Table 1: *Choi's features*

Discourse function	Topic	Contr Focus	Tail	Compl/Pres Focus
PROM	+	+	−	−
NEW	−	+	−	+

Table 2: *Butt and King's features*

Discourse function	Topic	Focus	Background	Completive Info
PROM	+	+	−	−
NEW	−	+	−	+

Butt & King (1996) adapted Choi's proposal, making the distinctions in Table 2. In Butt & King (1996) the PROM feature is not used to distinguish between contrastive and non-contrastive focus, as they only discuss cases of what they consider non-contrastive focus in their paper. This, of course, leaves open the question of which distinction needs to be made to account for the cases that have been discussed as contrastive versus non-contrastive focus.

Another difference between the two seems to be whether the IS structure should be a full representation of everything in the sentence or just some important parts. Whereas the Choi version of the features suggests that only some parts of the sentence will be represented, the Butt & King (1996) labeling assumes a full representation of the sentence. This latter view is also espoused in the more formal treatments (see Section 2.2).

Another version of the scheme is found in Gazdik & Komlósy (2011), who use the d-link distinction of Pesetsky (2007) instead of the ±NEW feature, as shown in Table 3. They consider continuing topics to be background, and discuss the difference between hocus and focus in Hungarian as well as the status of question words. This more recent proposal takes into account the most important distinctions discussed in the literature reviewed above. The notion of prominence seems to correspond to a notion of contrast, and d-linking is a way to distinguish between more and less salient elements. This proposal is, however, only worked out for Hungarian.

Table 3: *Gazdik and Komlosy features*

Discourse function	Focus, Question word, Hocus	Contrastive Topic, Question word	Completive	Background
PROM	+	+	−	−
D-LINKED	−	+	−	+

The switch in interpretation between Choi (1996) and Butt & King (1996) and the further switch in interpretation in Gazdik & Komlósy (2011) shows that the features are not clearly enough defined to apply unambiguously. It might be just this vagueness that has allowed several other LFG accounts, e.g Marfo & Bodomo (2005), Dalrymple & Nikolaeva (2011), Mycock (2013), Mycock & Lowe (2013) and Otoguro & Snijders (2016) to adopt the Choi (1996)/Butt & King (1996) approach.

What these two-feature approaches suggest is that, on an abstract level, only four distinctions need to be made to account for the IS distinctions that natural

languages encode, even if these distinctions are not exactly the same in all languages. This is not necessarily false but it is not something that has been argued for in any detail.

With respect to TOPIC, we often find one further distinction although some authors have proposed more subdivisions. Bresnan & Mchombo (1987) distinguish between CONTRASTIVE (new) topics and NON-CONTRASTIVE ones. Although they refer mainly to early work by Lambrecht, their distinction seems to be basically the Vallduví distinction between LINK and TAIL. The distinction between LINK and TAIL is also appealed to in Dalrymple & Nikolaeva (2011) as closely corresponding to theirs between PRIMARY and SECONDARY topics. Based on an analysis and data treated in more detail in Nikolaeva (2000), Dalrymple & Nikolaeva (2005) argue explicitly for a distinction between PRIMARY and SECONDARY TOPIC in Ostyak: according to the *what about X?* test, the primary topic has to be a subject in this language but agreeing objects are secondary topics. They are typical answers to questions such as *What did X*(= primary topic) *do to Y*(= secondary topic)? This analysis is further developed for several other languages in Dalrymple & Nikolaeva (2011) in the context of a discussion of DOM phenomena.

Abubakari's (2018) FAMILIARITY and CONTRASTIVE topics (in Kusaal) seem to be intended to capture the distinctions between SWITCH and CONTINUING topic but he seems to assume there could be more that two varieties of topic. Mchombo et al. (2005) make a distinction between ±CONTRASTIVE that they see as a further distinction within SWITCH TOPICS that are distinguished from CONTINUING TOPICS. Kifle (2011) proposes three topics in certain sentences in Tigrinya. Szűcs (2014) sees the distinction between CONTRASTIVE and NON-CONTRASTIVE topics (= −NEW elements) as crucial for left-dislocation and 'topicalization' in English: the topic position can be occupied by a contrastive element, be it topic or focus, whereas left-dislocation requires a non-contrastive new element. His distinctions seem to be similar to the ones made in Mchombo et al. (2005) but similarities or dissimilaries are not discussed.

Early work on FOCUS often distinguishes between CONTRASTIVE and PRESENTATIONAL focus (e.g. King (1995). A distinction between CONTRASTIVE and NON-CONTRASTIVE FOCI is made in Abubakari (2018). Dahlstrom (2003) appeals to Lambrecht's (1994) three-way distinction among foci: PREDICATE FOCUS, ARGUMENT FOCUS, SENTENCE FOCUS. Gazdik & Komlósy (2011) also distinguish between HOCUS and focus in Hungarian.

When four distinctions are felt not to be enough, various hierarchies are appealed to make further distinctions. The GIVENNESS HIERARCHY is invoked to make distinctions among topical and/or focal elements: see e.g. O'Connor (2006),

Andréasson (2008), Andréasson (2009), Andréasson (2013), who appeal to the notion ±ACTV (activated), Andreasson and Connor base themselves on the Gundel hierarchy but similar ideas are found in Lambrecht (1994).

Morimoto (2000) appeals to an animacy hierarchy in her analysis of subject-object inversion, and hence assumes that it plays a role in the assignment of the IS function subject in Bantu. Other hierarchies proposed are the Silverstein hierarchy (e.g. Simpson 2012, and similar hierarchies of topic-worthiness: Dalrymple & Nikolaeva 2011) or appeals to animacy, definiteness and specificity as in Mayer (2006). Mycock (2013) adds a feature for questions that can co-occur with all others. O'Connor (2006) add the feature ±OPEN to capture representations with and without a variable. The most extensive feature taxonomy in LFG, to my knowledge, has been proposed in Cook & Payne (2006) (see Section 3).

The LFG analyses as a whole would profit from more cross referencing and more discussion of the similarities and dissimilarities among the various proposals. An exception is Dalrymple & Nikolaeva (2011), who adopt the two-feature scheme of Butt & King (1996) and discuss how their notions of PRIMARY and SECONDARY TOPIC are different. The feature CONTRASTIVE is liberally used by various authors but often not further defined. Given how problematic it is, it would profit from a systematic clarification.

In general, not much attention is spent on the question of how to identify topic or focus independently of their syntactic, prosodic or morphological characteristics. So it is not always clear that the marking that is thought to be that of an IS unit might not mark another distinction.

2.2 Representation of IS in LFG

2.2.1 From f-structure functions to a separate IS representation

The first mentions of IS notions in LFG are the TOPIC/FOCUS function in Kaplan & Bresnan (1982) and Zaenen (1985). These are taken over from the phrase structure treatment of long distance dependencies in the grammatical frameworks that were then current. The discussion of what these discourse functions did was limited to the observation that they were "overlay" functions, requiring an extension of the coherence principle: topics or foci were not only topics or foci, but also had an argument function such as subject or object. The actual content of the notions topic and focus was not discussed. The discourse functions were treated in the f-structure, just like other functions. In the early nineties, several Stanford theses (e.g. Alsagoff 1992, Joshi 1993, Kroeger 1991) investigated the relation between SUBJECT and TOPIC in Asian languages in syntactic terms.

In the first studies that do discuss IS as such, the grammaticalized discourse function approach is also used, e.g. in Bresnan & Mchombo (1987) and in King (1995), who investigates in detail the phrase structure configuration needed to account for the configurational encoding of Russian discourse relations.

King (1997) discusses the drawbacks of an approach that integrates IS notions into the c-structure and the f-structure. She illustrates in detail the mismatches between f-structure units and IS units and proposes to handle IS as a separate projection. This is what most researchers have done in subsequent work. We will refer to this separate module as the *i-structure*.

As already indicated above, most researchers start from a two-feature analysis of topic and focus, in most cases augmented with background and completive roles. The representation given is generally an AVM, with the roles as attributes. The nature of the values depends on the way the relation of the i-structure to the other projections is articulated.

2.2.2 How does the IS relate to the other components of the grammar?

As IS can be signaled in various ways, the flow of information from the different components to the separate i-structure has to be modeled. LFG has a modular structure which allows researchers to experiment with various approaches while keeping other aspects of the framework constant. Overall, two models have been adopted by several researchers. One is proposed in the early paper by Butt & King (1996) and King (1997) and discussed further in Butt & King (1997); a later, different one, is proposed in Dalrymple & Nikolaeva (2011) and further work in glue semantics.

In Butt & King (1996) and Butt & King (1997), the c-structure feeds into the i-structure. The i-structure and the f-structure feed into the semantic structure and the i-structure is related to the f-structure, as every PRED appearing in the f-structure has to be linked to a discourse function. The BK model is assumed in Sulger (2009), Dione (2012) and Andréasson (2007), but it has not been worked out in detail.

Dalrymple & Nikolaeva (2011) develop a structured meaning approach à la von Stechow (1982) and Krifka (2006). In their view, the semantic structure encodes how meaning constructors relate to each other. The i-structure adds further structure specifying the pragmatic relations. Every meaning constructor in a sentence has to have a role at i-structure. What this role is can be positionally determined, through a c-structure annotation, or morphologically or prosodically. The feeding relations are c-structure to f-structure to s-structure to i-structure. The DN model is worked out in detail in Dalrymple et al. (2019). For discussions of how

the prosodic information fits in, see also Dalrymple & Mycock (2011), Mycock & Lowe (2013), Mycock (2013) and Bögel forthcoming [this volume]. Apart from these two proposals, there are proposals by individual researchers that draw attention to specific problems; e.g. O'Connor (2006) stresses the importance of the i-structure (his d-structure) relation to prosody. For him, part of the goal is to link an AVM representation for i-structure (his discourse structure) to a tree representation for prosody. Otoguro (2003) discusses the relation to morphology. Dahlstrom (2003) draws attention to the necessity of allowing constructional information to distinguish the various types of focus, especially sentence focus, and tentatively proposes an i-structure organized as a set of propositions.

Several researchers (e.g. O'Connor 2006, Choi 1996, Andréasson 2010) propose an Optimality-Theoretic calculation to determine what is topic or focus or what is reanalyzed in a particular way, but no precise proposals are made about how this OT part fits in with the rest of the architecture.

3 Studies of IS phenomena in LFG

In what follows, I list LFG contributions in IS in chronological order with some short comments intended to inform the reader which language data can be found in the contribution and which issues are most prominent.

Bresnan & Mchombo (1987) *Topic, pronoun, and agreement in Chicheŵa.* This early paper treats the IS concepts as part of the f-structure. It discusses mainly word order and the notions subject and object in Chicheŵa and some other Bantu languages.

King (1995) *Configuring topic and focus in Russian.* A revised version of a PhD thesis. Discusses topic, contrastive and presentational focus and background in Russian, Serbo-Croatian and Bulgarian. The IS notions are encoded in the c- and the f-structure, and other possible architectures are discussed.

Choi (1996) *Optimising structure in context: scrambling and information structure.* This PhD thesis discusses scrambling in German and Korean and appeals to the notions of aboutness topic and contrastive and presentational focus. It influenced later research by introducing the feature decomposition in ±NEW and ±PROM(inent) and its use of Optimality Theory to calculate the results.

Butt & King (1996) *Structural topic and focus without movement.* The paper discusses word order and discourse configurationality in Urdu and Turkish and distinguishes topic, focus, background and completive information. Influential in the new way it used the features ±NEW, ±PROM.

King (1997) *Focus domains and information-structure.* The paper explicitly discusses the problem created by representing IS in the c- and the f-structure on the basis of Russian data. It proposes a i-structure parallel to the f-structure.

Sharma (1999) *Nominal clitics and constructive morphology in Hindi.* The focus of this paper is the representation of focus clitics in Hindi via inside-out uncertainty.

Broadwell (1999) *The interaction of focus and constituent order in San Dionicio Zapotec.* The paper uses Optimality Theory to calculate the right word order for focused constituents in Zapotec.

Morimoto (2000) *Discourse configurality in Bantu morphosyntax*, see also Morimoto (2009). This dissertation looks at Kirundi and Kinyarwanda and discusses subject-object inversion. Following Bresnan (2001) in analysing SUBJECT as both an argument and a discourse function and using the features of Choi (1996), it argues for two notions of topic in Bantu: external and internal topic. The distinctions are encoded in the f-structure.

Butt & King (1997) *Null elements in discourse structure.* The paper discusses pro-drop in Hindi/Urdu. It adds the distinction between switch and continuing topic to the distinctions made in Butt & King (1996). It uses a separate i-structure projected mainly from the c-structure.

Otoguro (2003) *Focus clitics and discourse information spreading*. The paper studies focus clitics in Japanese and argues for an architecture in which the c-structure is the input to the i-structure as well as to the f-structure and both are input to the morphology. It uses Optimality Theory to calculate the outcomes.

Dahlstrom (2003) *Focus constructions in Meskwaki.* The paper starts from Lambrecht's (1994) three focus types and discusses the various constructions, exemplifying them in Meskwaki. Following Lambrecht (1994), it proposes an i-structure that is structured as a set of propositions.

Dalrymple & Nikolaeva (2005) *Non-subject agreement and discourse roles.* The paper makes an argument for the notion secondary topic based on agreement facts in Ostyak. It assumes a separate i-structure.

Marfo & Bodomo (2005) *Information structuring in Akan question word fronting and focus constructions.* Starting from Akan question word fronting, the paper studies the difference between focus and background. It assumes a separate i-structure.

Mchombo et al. (2005) *Partitioning discourse information: a case of Chicheŵa split constituents.* The paper argues that the topic in Chicheŵa can be split into a −PROM and a +PROM part. The +PROMINENT part can be in initial position or not and it can be ±CONTRASTIVE.

Cook & Payne (2006) *Information structure and scope in German.* The paper examines the interaction between word order and scope in German and claims we need to distinguish between ±TOPIC, ±NEW and ±CONTRASTIVE. The facts discussed relate to what others have called topic-within-focus. The account uses a separate i-structure, glue semantics and Optimality Theory.

Mayer (2006) *Optional direct object clitic doubling in Limeño Spanish.* The paper discusses clitic doubling in Limeño Spanish. It contains an extensive discussion of the factors that are usually associated with differential object marking (DOM): animacy, definiteness and specificity. It proposes that some of the objects discussed might be secondary topics.

O'Connor (2006) *Information structure in lexical-functional grammar: the discourse-prosody correspondence.* This dissertation discusses prosody and pitch accent in Serbo-Croatian and their link to IS notions. It is based on Lambrecht's (1994) distinction between presupposition and assertion and the distinction between active and nonactive referents. In O'Connor's terminology, discourse structure corresponds to what is called i-structure in this paper. He represents discourse structure as an AVM and discusses how it should be linked to the prosodic structure that is represented as a tree.

Simpson (2007) *Expressing pragmatic constraints on word order in Warlpiri.* The paper discusses word order in Warlpiri. It argues for a distinction between prominent and nonprominent information as well as the distinction between new and not new. The AUX marks the transition from prominent to less prominent information. New information precedes the verb. Both prominence and newness are seen as relational notions. The separate IS is intended to be capable of representing hierarchies of newness and prominence.

Kifle (2007) *Differential object marking and topicality in Tigrinya.* see also Kifle (2011). The discussion is based on Dalrymple & Nikolaeva (2005) but it is claimed that for Tigrinya further distinctions are needed. The topic is represented at f-structure in the implementation.

Andréasson (2007) *The architecture of i-structure.* The paper argues for a function SCENE distinct from GROUND and RHEME and from STAGE TOPIC. The data come from Scandinavian languages, mainly Swedish, where the SCENE is place

between the GROUND and the RHEME. A separate i-structure is assumed but it is not clear how it relates to the rest of the grammar.

Andréasson (2008) *Not all objects are born alike.*
Andréasson (2009) *Pronominal object shift — not just a matter of shifting or not.*
Andréasson (2010), *Object shift or object placement in general?*
Andréasson (2013) *Object shift in Scandinavian languages: the impact of contrasted elements.* This series of papers studies the different factors that influence Object Shift in Scandinavian, especially in Danish and Swedish. They argue for an accessibility hierarchy a la Gundel et al. (1993) (Andréasson 2008), the importance of factivity when clausal antecedents are involved (Andréasson 2009) and the role of contrastive focus (Andréasson 2013). They assume a separate i-structure and discuss its link to the c-structure.

Sulger (2009) *Irish clefting and information-structure.* The paper argues for the distinction between GROUND and FOCUS in Irish clefts. It assumes a separate i-structure projected from the c-structure.

Gazdik (2010) *Multiple questions in French and in Hungarian. An LFG account.* The paper studies questions in French and in Hungarian, making a distinction between FOCUS, TOPIC and BACKGROUND using a separate i-structure.

Dalrymple & Nikolaeva (2011) *Objects and information structure.* This book discusses (differential) object marking and agreement in several languages (Uralic (Ostyak, Tundra Nenets, Vogul), Iranian languages, Indo-Aryan) with a typological and historical perspective. It mainly discusses primary and secondary topics and distinguishes the notion of topic from that of topic-worthiness which is based on prominence features such as animacy, definiteness, and specificity. It uses Lambrecht's (1994) notions of assertion and presupposition. It proposes a separate i-structure and provides a structured meaning representation for topic and focus projected from the semantic structure.

Gazdik & Komlósy (2011) *On the syntax-discourse interface in Hungarian.* The paper discusses word order and prosody in Hungarian preverbal field. It distinguishes between the hocus (an element that highlights an unusual feature of a otherwise usual event) and the focus and proposes a revision of the Butt & King (1996) schema appealing to the notion of d-linking. It uses a separate i-structure.

Simpson (2012) *Information structure, variation and the referential hierarchy.* The paper discusses agreement and word order in Warlpiri and Arrernte and points out the importance of the Silverstein hierarchy to account for the data. It doesn't address architectural issues.

Dione (2012) *An LFG approach to Wolof cleft constructions.* The paper mainly discusses clefts in Wolof. It argues that the i-structure can be part of the f-structure when it has been syntactized.

Mycock (2013) *Discourse functions of question words.* The paper discusses questions in English and Urdu/Hindi. It follows the Butt & King (1996) proposal but adds a Q mark to all distinctions. It assumes a separate i-structure.

Mycock & Lowe (2013) *The prosodic marking of discourse functions.* The paper discusses the prosody of broad and narrow focus in English. The IS distinctions are based on Dalrymple & Nikolaeva (2011). It addresses the relation between c-structure, i-structure and p-structure.

Butt (2014) *Questions and information structure in Urdu/Hindi.* The paper develops the distinctions made in Butt & King (1996) proposing more subdivisions to account for questions in Urdu/Hindi. It assumes a separate i-structure but doesn't discuss the relation between projections.

Szűcs (2014) *Information structure and the English left periphery.*
Szűcs (2017) *English left-peripheral constructions from an LFG perspective.* These papers discuss the English left periphery based on insights from Prince (1981) and Ward & Birner (2001). They argue for a distinction between ±NEW and ±D-LINKED which is further subdivided into ±CONTRASTIVE. The IS notions are represented in the f- and the c-structure.

Zymla et al. (2015) *Modeling the common ground for discourse particles.* The paper discusses discourse particles in German in the context of the PARGRAM AKR (Abstract Knowledge Representation).

Otoguro & Snijders (2016) *Focus clitics and discourse information spreading.* The paper discusses quantifier float in Dutch, English and Japanese. Based on the Butt & King (1996) distinctions it argues that quantified NPs are topics and the floated quantifer is part of the focus.

Belyaev (2017) *Information structure conditions on agreement controller in Dargwa.* The paper argues for the importance of the notion PIVOT as defined in Falk (2006) to account for agreement in Dargwa. The notions used are syntactically encoded.

Abubakari (2018) *Information structure and the Lexical-Functional framework.* The paper argues for a subdivision of focus in contrastive and completive based on data from the morphological markings on focus and topic in Kusaal. The morphological markers themselves are retained in the i-structure.

Szűcs (2019) *Left-dislocation in Hungarian.* The paper argues for a distinction between topic left dislocation and clitic topicalization in Hungarian. It mainly discusses the f- and the c-structure.

Acknowledgements

Thanks to three anonymous reviewers and to Tracy Holloway King for comments on an earlier version. Needless to say, only I am responsible for any remaining errors or omissions.

References

Abubakari, Hasiyatu. 2018. Information structure and the Lexical-Functional framework. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '18 conference*, 4–24. Stanford, CA: CSLI Publications.

Alsagoff, Lubna. 1992. *Topic in Malay: The other subject.* Stanford, CA: Stanford University. (Doctoral dissertation).

Andréasson, Maia. 2007. The architecture of i-structure. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '07 conference*, 26–43. Stanford, CA: CSLI Publications.

Andréasson, Maia. 2008. Not all objects are born alike. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 26–45. Stanford, CA: CSLI Publications.

Andréasson, Maia. 2009. Pronominal object shift — Not just a matter of shifting or not. *Working Papers in Scandinavian Syntax* 84.

Andréasson, Maia. 2010. Object shift or object placement in general? In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 26–42. Stanford, CA: CSLI Publications.

Andréasson, Maia. 2013. Object shift in Scandinavian languages: The impact of contrasted elements. *Nordic Journal of Linguistics* 36. DOI: 10.1017/s0332586513000231.

Ariel, Miriam. 1988. Referring and accessibility. *Journal of Linguistics* 24. DOI: 10.1017/s0022226700011567.

Belyaev, Oleg. 2017. Information structure conditions on agreement controller in Dargwa. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '17 conference*, 66–82. Stanford, CA: CSLI Publications.

Bentley, Delia & Silvio Cruschina. 2018. The silent argument of broad focus: Typology and predictions. *Glossa: A Journal of General Linguistics* 3. 1–17. DOI: 10.5334/gjgl.677.

Bickel, Balthasar. 2012. On the scope of the referential hierarchy in the typology of grammatical relations. In Greville G. Corbett & Michael Noonan (eds.), *Case and grammatical relations: Studies in honor of Bernard Comrie*. Amsterdam: John Benjamins.

Bögel, Tina. Forthcoming. Prosody and its interfaces. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

Booth, Hannah & Miriam Butt. Forthcoming. LFG and historical linguistics. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

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.

Broadwell, George Aaron. 1999. The interaction of focus and constituent order in San Dionicio Zapotec. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '99 conference*. Stanford, CA: CSLI Publications.

Büring, Daniel. 2003. On d-trees, beans and b-accents. *Linguistics and Philosophy* 26.

Butt, Miriam. 2014. Questions and information structure in Urdu/Hindi. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 158–178. Stanford, CA: CSLI Publications.

Butt, Miriam & Tracy Holloway King. 1996. Structural topic and focus without movement. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '96 conference*. Stanford, CA: CSLI Publications.

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.

Chafe, Wallace L. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li (ed.), *Subject and topic*, 27–55. New York: Academic Press.

Chafe, Wallace L. 2007. Cognitive constraints on information flow. In Russell S. Tomlin (ed.), *Coherence and grounding in discourse*, 21–51. Amsterdam: John Benjamins.

Choi, Hye-Won. 1996. *Optimising structure in context: Scrambling and information structure*. Stanford, CA: Stanford University. (Doctoral dissertation).

Cook, Philippa & John R. Payne. 2006. Information structure and scope in German. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '06 conference*. Stanford, CA: CSLI Publications.

Dahl, Östen. 1974. Topic-comment structure revisited. In Östen Dahl (ed.), *Topic and comment, contextual boundedness and focus*. Helmut Buske.

Dahlstrom, Amy. 2003. Focus constructions in Meskwaki. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '03 conference*, 144–163. 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. 2005. Nonsubject agreement and discourse roles. In Anna McNay (ed.), *Oxford working papers in linguistics, philology, and phonetics*. Oxford: Centre for Linguistics & Philology, University of Oxford.

Dalrymple, Mary & Irina Nikolaeva. 2011. *Objects and information structure* (Cambridge Studies in Linguistics). Cambridge, UK: Cambridge University Press.

Dik, Simon et al. 2008. On the typology of focus phenomena. In Teun Hoekstra, Harry van der Hulst & Michael Moortgat (eds.), *Perspectives on functional grammar*. Dordrecht: Foris Publications.

Dione, Cheikh M. Bamba. 2012. An LFG approach to Wolof cleft constructions. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 157–176. Stanford, CA: CSLI Publications.

Dixon, R. M. W. 1994. *Ergativity*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/cbo9780511611896.

Dryer, Matthew S. 1995. Focus, pragmatic presupposition and activated propositions. *Journal of Pragmatics* 26. DOI: 10.1016/0378-2166(95)00059-3.

Endriss, Cornelia. 2009. *Quantificational topics: A scopal treatment of exceptional wide scope phenomena*. Springer. DOI: 10.1007/978-90-481-2303-2.

Erteschik-Shir, Nomi. 2007. *Information structure: The syntax-discourse interface*. Oxford: Oxford University Press.

Falk, Yehuda N. 2006. *Subjects and universal grammar: An explanatory theory*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/cbo9780511486265.

Firbas, Jan. 1975. On the thematic and the non-thematic section of the sentence. In Håkan Ringbom et al. (eds.), *Style and text: Studies presented to Nils Erik Enkvist*. Stockholm.

Frascarelli, Mara & Roland Hinterhölzl. 2007. Types of topics in German and Italian. In Kerstin Schwabe & Susanne Winkler (eds.), *On information structure, meaning and form*, 87–116. Amsterdam: John Benjamins. DOI: 10.1075/la.100.07fra.

Frey, Werner. 2006. Contrast and movement to the German prefield. In Valéria Molnár & Susanne Winkler (eds.), *The architecture of focus*. Berlin: Mouton de Gruyter.

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 & 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.

Gécseg, Zsuzsanna & Ferenc Kiefer. 2009. A new look at information structure in Hungarian. *Natural Language & Linguistic Theory* 27. 583–622. DOI: 10.1007/s11049-009-9071-7.

Givón, Talmy. 1983. Introduction. In Talmy Givón (ed.), *Topic continuity in discourse*. Amsterdam: John Benjamins.

Gundel, Jeanette. 1974. *The role of topic and comment in linguistic theory*. Austin: University of Texas at Austin. (Doctoral dissertation).

Gundel, Jeanette, Nancy Hedberg & Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2). DOI: 10.2307/416535.

Halliday, M. A. K. 1967. Note on transitivity and theme in English. *Journal of Linguistics* 3.

Hamblin, Charles. 1973. Questions in Montague grammar. *Foundations of Language* 10. 41–53.

Hedberg, Nancy. 2006. Topic-focus controversies. In Valéria Molnár & Susanne Winkler (eds.), *The architecture of focus*. Berlin: Mouton de Gruyter.

Heim, Irene. 1982. *The semantics of definite and indefinite noun phrases*. Amherst: University of Massachusetts, Amherst. (Doctoral dissertation).

Jackendoff, Ray. 1972. *Semantic interpretation in generative grammar*. Cambridge, MA: The MIT Press.

Jacobs, Joachim. 1984. Functionale Satzperspektive and Illukutionssemantik. *Linguistische Beriche* 91.

Joshi, Smita. 1993. *Selection of grammatical and logical functions in Marathi*. Stanford, CA: Stanford University. (Doctoral dissertation).

Kamp, Hans. 1981. A theory of truth and semantic representation. In Jeroen Groenendijk, Theo M. V. Janssen & Martin Stokhof (eds.), *Formal methods in the study of language*. Mathematics Center, Amsterdam. DOI: 10.1002/9780470758335.ch8.

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).

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.

King, Tracy Holloway. 1995. *Configuring topic and focus in Russian*. Stanford, CA: CSLI Publications.

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.

King, Tracy Holloway & Annie Zaenen. 2004. F-structures, information structure, and discourse structure. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*. Stanford, CA: CSLI Publications.

É. Kiss, Katalin. 1998. Identificational focus versus information focus. *Language* 74(2). DOI: 10.2307/417867.

Krifka, Manfred. 2006. Association with focus phrases. In Valéria Molnár & Susanne Winkler (eds.), *The architecture of focus*. Berlin: Mouton de Gruyter.

Krifka, Manfred. 2007. The notions of information structure. In Caroline Féry, Gisbert Fanselow & Manfred Krifka (eds.), *Interdisciplinary studies on information structure*. Potsdam.

Kroeger, Paul R. 1991. *Phrase structure and grammatical relations in Tagalog*. Stanford, CA: Stanford University. (Doctoral dissertation).

Kruijff-Korbayová, Ivana & Mark Steedman. 2003. Discourse and information structure. *Journal of Logic, Language, and Information* 12(3). 249–259.

Kuno, Susumu. 1987. *Functional syntax*. Chicago: University of Chicago Press.

Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representation of discourse*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/cbo9780511620607.

Marfo, Charles O. & Adams Bodomo. 2005. Information structuring in Akan question-word fronting and focus constructions. *Studies in African Linguistics* 34. DOI: 10.32473/sal.v34i2.107330.

Matič, Dejan & Daniel Wedgwood. 2013. The meanings of focus: The significance of an interpretation-based category in cross-linguistic analysis. *Journal of Linguistics* 49. DOI: 10.1017/S0022226712000345.

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.

Mchombo, Sam A., Yukiko Morimoto & Caroline Féry. 2005. Partitioning discourse information: A case of Chichewa split constituents. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '05 conference*. Stanford, CA: CSLI Publications.

Morimoto, Yukiko. 2000. *Discourse configurality in Bantu morphosyntax*. Stanford, CA: Stanford University. (Doctoral dissertation).

Morimoto, Yukiko. 2009. From topic to subject marking: Implications for a typology of subject marking. In Helen de Hoop & Peter de Swart (eds.), *Differential subject marking*. Springer. DOI: 10.1007/978-1-4020-6497-5_9.

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.

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.

Neeleman, Ad & Reiko Vermeulen. 2013. The syntactic expression of information structure. In Ad Neeleman & Reiko Vermeulen (eds.), *Syntax of topic, focus and contrast*. Berlin: De Gruyter.

Nikolaeva, Irina. 2000. Secondary topic as a relation in information structure. *Linguistics* 39. 1–49. DOI: 10.1515/ling.2001.006.

O'Connor, Robert. 2006. *Information structure in Lexical-Functional Grammar: The discourse-prosody correspondence*. Manchester: University of Manchester. (Doctoral dissertation).

Otoguro, Ryo. 2003. Focus clitics and discourse information spreading. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '03 conference*. Stanford, CA: CSLI Publications.

Otoguro, Ryo & Liselotte Snijders. 2016. Syntactic, semantic and information structures of floating quantifiers. 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*. Stanford, CA: CSLI Publications.

Pesetsky, David. 2007. *Wh*-in-situ: Movement and unselective binding. In Eric J. Reuland & Alice G. B. ter Meulen (eds.), *The representation of (in)definiteness*. Cambridge, MA: The MIT Press.

Prince, Ellen. 1981. Towards a taxonomy of given/new information. In Peter Cole (ed.), *Radical pragmatics*, 223–255. Academic Press.

Reinhart, Tanya. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica* 27.

Repp, Sophie. 2016. Contrast: Dissecting an elusive information-structural notion and its role in grammar. In Caroline Féry & Shinichiro Ishihara (eds.), *The Oxford handbook of information structure*. Oxford: Oxford University Press.

Rooth, Mats. 1992. A theory of focus interpretation. *Natural Language Semantics* 1. 75–116. DOI: 10.1007/bf02342617.

Sharma, Devyani. 1999. Nominal clitics and constructive morphology in Hindi. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '99 conference*. Stanford, CA: CSLI Publications.

Silverstein, Michael. 1976. Hierarchy of features and ergativity. In R. M. W. Dixon (ed.), *Grammatical categories in Australian languages*. Australian Institute of Aboriginal Studies. DOI: 10.1515/9783110871661-008.

Simpson, Jane. 2007. Expressing pragmatic constraints on word order in Warlpiri. 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*. Stanford, CA: CSLI Publications.

Simpson, Jane. 2012. Information structure, variation and the referential hierarchy. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts & Paul Trilsbeek (eds.), *Potentials of language documentation: Methods, analyses, and utilization*, 83–89. Honolulu: University of Hawai'i Press.

Stalnaker, Robert. 1970. Pragmatics. *Synthese* 22. DOI: 10.1007/978-94-010-2557-7_11.

Steedman, Mark. 2000. Information structure and the syntax-phonology interface. *Linguistic Inquiry* 31(4). 649–689. DOI: 10.1162/002438900554505.

Strawson, Peter. 1964. Identifying reference and truth values. *Theoria* 30. 96–118. DOI: 10.1111/j.1755-2567.1964.tb00404.x.

Sulger, Sebastian. 2009. Irish clefting and information-structure. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 562–582. Stanford, CA: CSLI Publications.

Szűcs, Péter. 2014. Information structure and the English left periphery. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '14 conference*, 525–565. Stanford, CA: CSLI Publications.

Szűcs, Péter. 2017. English left-peripheral constructions from an LFG perspective. *Linguistica Brunensia* 65.

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.

Vallduví, Enric. 1992. *The informational component*. New York: Garland.

Vallduví, Enric & Elisabet Engdahl. 1996. The linguistic realization of information packaging. *Linguistics* 34. DOI: 10.1515/ling.1996.34.3.459.

Vallduví, Enric & Maria Vilkuna. 1998. On rheme and kontrast. In Peter Culicover & Louise McNally (eds.), *The limits of syntax* (Syntax and Semantics 29). Academic Press. DOI: 10.1163/9789004373167_005.

von Stechow, Arnim. 1982. *Structured propositions*. Universität Konstanz SFB.

Ward, Gregory & Betty Birner. 2001. Discourse and information structure. In Deborah Tannen, Heidi E. Hamilton & Deborah Schiffrin (eds.), *The Oxford handbook of discourse analysis*. Oxford: Oxford University Press. DOI: 10.1002/9780470753460.ch7.

Zaenen, Annie. 1985. *Extraction rules in Icelandic*. New York: Garland.

Zaenen, Annie. Forthcoming. Information structure. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

Zymla, Mark-Matthias, Maike Müller & Miriam Butt. 2015. Modeling the common ground for discourse particles. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '15 conference*, 420–440. Stanford, CA: CSLI Publications.

Chapter 2

LFG treebanks

Victoria Rosén
University of Bergen

Treebanks are syntactically annotated corpora. LFG treebanks are collections of LFG analyses, usually created by parsing a corpus with an LFG grammar. This chapter provides an overview of existing LFG treebanks, explains how they are created, how they may be searched, and what their potential use is to the LFG and other communities.

1 Introduction

Annotated corpora are important resources for many branches of linguistics, language studies and natural language processing. A common form of corpus annotation consists of labeling words with their parts of speech, lemmas and morphosyntactic features, such as number, person, tense, etc. Using only annotation at the word level limits the potential to search for important grammatical information, such as syntactic constructions, grammatical functions and predicate–argument relations. The usefulness of corpora is therefore greatly enhanced if they also include syntactic annotation, such as phrase structure and functional relations. Syntactically annotated corpora are usually called treebanks; if they are created by parsing, they may also be called parsed corpora or parsebanks.

LFG treebanks are treebanks annotated according to the LFG formalism. They are usually created as parsebanks, by parsing a corpus with an LFG grammar and disambiguating the parse results. An LFG parsebank is thus essentially a collection of analyses according to a grammar. LFG parsebanks encode a wealth of morphological, syntactic and semantic information in their c- and f-structure representations, and tend to be more detailed than treebanks adhering to other

Victoria Rosén. Forthcoming. LFG treebanks. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press. 

formalisms. The term treebank is well established even if the treebank may contain f-structures, which are directed graphs rather than trees.

This chapter is aimed at two audiences. The first target group consists of linguists who may wish to learn to use LFG treebanks in order to find data for their research. The second target group is linguists who may wish to build LFG treebanks as part of a grammar development project.

A major platform for LFG treebanking is INESS (Infrastructure for the Exploration of Syntax and Semantics) at the CLARINO Bergen Center (University of Bergen, Norway).[1] This infrastructure will be further introduced below and will be used throughout the chapter to illustrate the various possibilities of LFG treebanking.

Section 2 describes how LFG treebanks can be created through parsing with the Xerox Linguistic Environment and further processed with the LFG Parsebanker. In Section 3 the LFG treebanks in the INESS treebanking infrastructure are presented. Section 4 demonstrates how LFG treebanks may be searched with INESS Search. Finally, Section 5 describes approaches to conversion between LFG treebanks and treebanks adhering to other formalisms.

2 Building LFG treebanks

2.1 Basic requirements

A parser, an implemented grammar and lexicon, and efficient disambiguation tools are prerequisites for creating a parsebank. A useful set of tools in this respect is the Xerox Linguistic Environment (XLE), developed at the Palo Alto Research Center and the Xerox Research Centre Europe in Grenoble. XLE includes both a parser and a generator for LFG grammars, and it is suitable for grammar implementation on a small or large scale (Crouch et al. 2008; Maxwell & Kaplan 1993). For detailed information on XLE, see Forst & King forthcoming [this volume].

A grammar and lexicon with wide coverage are essential for building a large treebank of authentic texts, as well as for other applications. Grammar development is however a process which typically starts with a small set of rules which is successively expanded. In this development, the grammar must constantly be tested to see whether all the old rules still work in addition to the new rules. In this incremental process, a corpus, even a small one initially, may be useful as a

[1] https://clarino.uib.no/iness. INESS was built with funding from the Research Council of Norway in the eponymous project (2010-2017) (Rosén et al. 2012; Meurer et al. 2013).

test suite for parsing. As the grammar grows, it can be tested on a larger corpus and further improved. Larger grammars and lexicons do however increase the ambiguity in the analyses, so that efficient disambiguation is important.

XLE-Web[2] is a web-based implementation of XLE that was first developed in the LOGON and TREPIL projects (Rosén et al. 2005, 2006). XLE-Web uses the same parsing technology and software as XLE, but differs from the original platform in several ways. The original XLE is a standalone, integrated platform for grammar writing and debugging, whereas XLE-Web can be used through any modern browser. XLE-Web does not have tools for grammar writing, but it has excellent tools for disambiguation.

As mentioned above, ambiguity becomes a considerable problem as the grammar grows. Therefore, XLE-Web offers *discriminant disambiguation* to efficiently select the intended analysis among possibly many alternative analyses. Discriminant analysis is a technique for identifying minimal differences between analyses and letting disambiguation proceed by resolving these differences rather than on inspection of whole structures (Rosén et al. 2007). An example of the XLE-Web display with discriminants is provided in Figure 1 for the ambiguous sentence *He saw the girl with binoculars*,[3] parsed with the English ParGram grammar.[4]

This sentence has two analyses due to a PP attachment ambiguity: *with binoculars* may be either an ADJUNCT of the clause or an ADJUNCT in the OBJ. Whereas XLE offers packed f-structures, XLE-Web offers packed representations for both c- and f-structures. A packed representation presents all analyses in one graph, with indices at choice points. In the middle of Figure 1 is a packed c-structure with one choice point which splits into the subtrees labeled *a1* and *a2*. A corresponding choice can be seen in the packed f-structure shown on the right in the figure. Although the disambiguated f-structure will have an ADJUNCT either on the outer level or inside the OBJ, both functions occur in the packed f-structure, labeled with *a1* and *a2* respectively.

On the left in the figure is a table with discriminants computed on the basis of these choice points. They present the user with each individual distinction between the analyses. There are two f-structure discriminants and ten c-structure discriminants.[5] F-structure discriminants describe paths through the f-structure

[2] https://clarino.uib.no/iness/xle-web

[3] In this and many subsequent f-structures, the PREDS only mode of display has been chosen. PREDS only mode displays only PRED values and the attribute paths which lead to them. This mode is often preferred when a full f-structure is too large to be easily legible.

[4] This grammar was developed in the Parallel Grammar (ParGram) project, see Section 3.6.2.

[5] In some cases there may also be lexical and morphological discriminants, but not for this sentence, which does not display any lexical ambiguities.

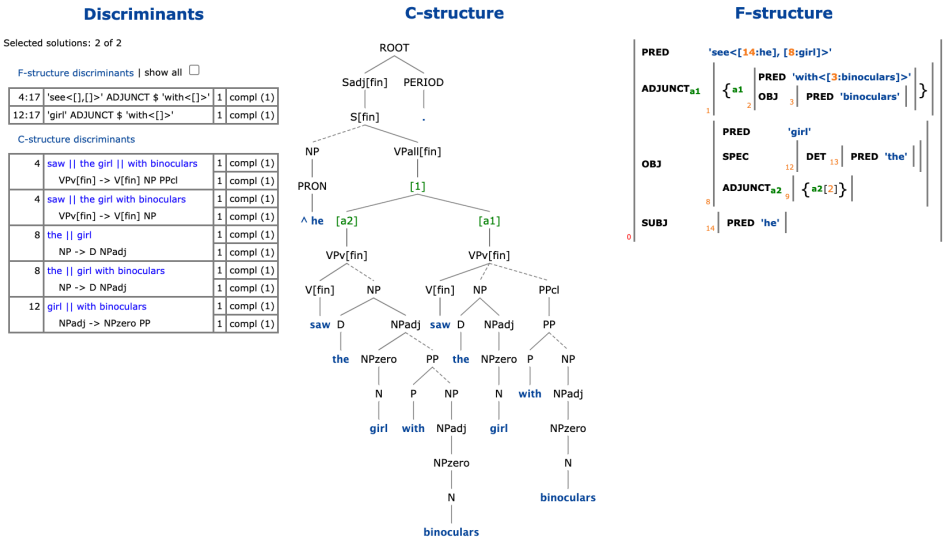


Figure 1: Analysis with discriminants and packed c- and f-structures for *He saw the girl with binoculars.*

from a PRED value to another PRED value or an atomic value. The two f-structure discriminants shown here indicate that the phrase *with binoculars* is an ADJUNCT either of the verb *see* or of the noun *girl*. The ten c-structure discriminants present the various minimal subtrees (a minimal subtree being a mother node and its daughter nodes) that make up the subtrees indexed with *a1* and *a2*. C-structure discriminants are either constituent discriminants, which show the bracketing of a substring, or rule discriminants, which show the labeled bracketing of a substring, expressed as a phrase structure rule. Rule discriminants are always displayed directly under the corresponding consituent discriminant, thus showing clearly which string of words the rule represents a bracketing of.

A discriminant may be chosen by clicking on it, or rejected by clicking on *compl* (for *complement*).[6] After a discriminant or its complement has been clicked on, it is displayed in boldface; the choice may be reversed by clicking on the boldfaced discriminant, thus resetting it. Since there are only two analyses for the

[6]The numbers to the left of the discriminants are anchors, which are necessary in case the same word or phrase occurs more than once in the sentence. In c-structure discriminants the anchor identifies the position of the first character in the substring. In f-structure discriminants the anchors identifiy the position of the first character of the words that project the PRED values in the discriminant. The number to the right of a discriminant or its complement indicates the number of solutions that will remain after it is chosen.

sentence in Figure 1, the intended one may be selected by choosing or rejecting any one discriminant. Figure 2 shows the effect of choosing the analysis in which *with binoculars* is an adjunct of the verb *see* by clicking on the first f-structure discriminant, resulting in full disambiguation. Discriminants that have not been chosen and that are no longer relevant for disambiguation, because they do not distinguish between any remaining analyses, are not displayed. This is important for efficiency, since the disambiguator then has fewer discriminants to take into consideration.

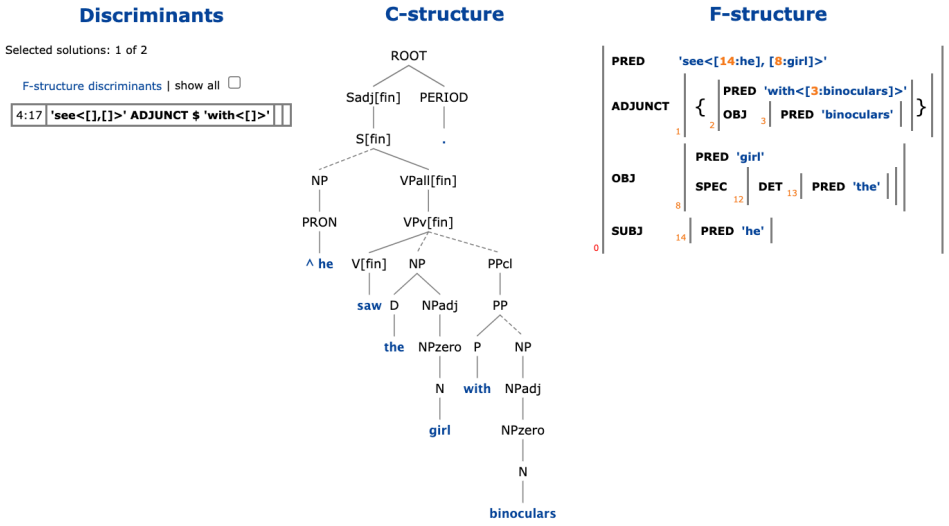


Figure 2: Fully disambiguated analysis for *He saw the girl with binoculars.*

This process may seem like overkill for this simple example which has only two readings. It becomes rewarding, however, when there are multiple ambiguities in the sentence. Even when the combination of ambiguities may give rise to a very large number of analyses, the number of discriminants does not necessarily increase as much, so that discriminant analysis remains comparatively efficient. A more detailed presentation of disambiguation with discriminants in LFG may be found in Rosén et al. (2007).

At the time of writing, the XLE-Web instance at INESS offers online parsing with the ParGram grammars of the following languages: English, French, Georgian, German, Indonesian, Italian, Malagasy, Norwegian, Polish, Tamil, Tigrinya, Turkish, Urdu and Wolof. Some of these have broad coverage, others are more limited in scope.

2.2 The LFG Parsebanker

The LFG Parsebanker, available in INESS, is an integrated set of tools for creating and searching LFG treebanks (Rosén et al. 2009). It allows texts to be batch parsed with the XLE parser, and it stores the analyses in a database. The resulting parsebank may be disambiguated by using discriminants in the same way as described above. The LFG Parsebanker stores both the analyses and all discriminant choices that were made. This means that the grammar and lexicon may be further developed, and the treebank subsequently reparsed and at least partially redisambiguated with the stored discriminant choices. This method makes it possible to develop the grammar and the treebank in tandem, thus incrementally improving the quality of the analyses. The stored discriminants may also be used for stochastic parseranking. In this way larger parsebanks can be automatically disambiguated.

A possible drawback of constructing a treebank by parsing with an LFG (or other) broad-coverage unification grammar is that the grammar cannot hope to have full coverage for all authentically occurring sentences in a large corpus. Nevertheless, some traditional treebanks that are (at least partially) manually annotated are meant to assign an analysis to every sentence, and a variety of methods are utilized to achieve this. When a sentence is not covered by the grammar, an annotator can, for instance, hand construct an analysis to "fix" the problem. Although this provides an analysis for the treebank, it does not provide an analysis that is consistent with a grammar, and sentences that are not actually grammatical may receive analyses as if they were. In contrast, a pure parsebank does not resort to such ad hoc fixes, since it is often primarily meant to test the coverage and precision of a grammar, so that it is desirable to keep the treebank in sync with the grammar. The LFG Parsebanker therefore does not permit disambiguators to edit the automatically derived analyses, but allows them to make notes for grammar and lexicon development to solve coverage problems.

3 LFG treebanks in INESS

INESS is a treebanking infrastructure for building, hosting and exploring treebanks. It includes the above-mentioned XLE-Web and the LFG Parsebanker. It also has an elaborate infrastructure for browsing, search and visualization, as will be explained below.

INESS accommodates not only LFG treebanks, but also treebanks based on other frameworks, such as HPSG, constituency, and dependency treebanks. The

infrastructure makes treebanks available online in an internet browser, eliminating the need to download treebanks and software for viewing and searching them, thus considerably facilitating access to them. Since INESS hosts many treebanks, there is an interface for treebank selection, as described in 3.1.

While some treebanks have completely open access, others require user authentication and authorization. Treebank owners decide under what licensing terms their treebanks are to be made available; some treebanks have restrictive licenses due to copyright of the input texts. The most open license that copyright will allow is recommended (Rosén & De Smedt 2022). INESS participates in the CLARIN Service Provider Federation (SPF), which allows researchers to authenticate themselves by logging in with their own university identity, thus gaining access to many more treebanks than are freely available. The CLARIN SPF has participant institutions in many countries, both in Europe and beyond. Users not belonging to one of these institutions can apply for a user name and password at CLARIN.[7]

INESS hosts LFG treebanks of varying sizes. The larger treebanks TIGER, the LFG Structure Bank for Polish, and NorGramBank are presented in 3.2, 3.3 and 3.4, respectively. The smaller treebanks are presented in 3.5. INESS also hosts several parallel treebanks with LFG annotations, presented in 3.6. The INESS interface is described in more detail by Meurer et al. (2020).

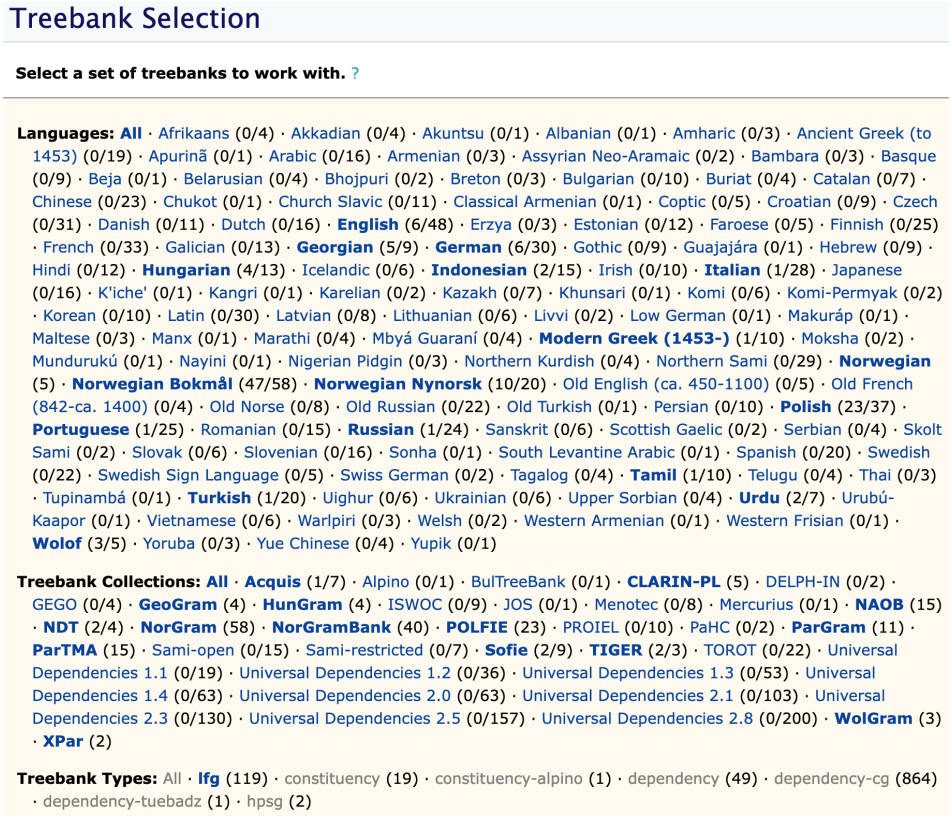
3.1 Selecting treebanks in INESS

The first step in exploring treebanks involves selecting one or more treebanks. At the time of writing, INESS hosts 433[8] treebanks for 115[9] languages. The *Treebank Selection* page in INESS, shown in Figure 3, groups treebanks according to language, collection and type.

[7]CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources (http://clarin.eu).

[8]According to Figure 3, there are 1057 treebanks in total, but this number includes all of the versions of the UD treebanks. If we only count the number of treebanks in Universal Dependencies 2.5 (200), the total number of treebanks is 433.

[9]There are 117 language names, but three of these are Norwegian, Norwegian Bokmål, and Norwegian Nynorsk, and these have been counted as one language: Norwegian. Norwegian Bokmål and Norwegian Nynorsk are the two written standards for the Norwegian language, with a good deal of lexical variation and many differences in spelling and morphology. Most treebank texts are written consistently in one variety or the other, so that users can choose which written variety to explore. Some texts, however, contain both varieties, for instance the proceedings of the Norwegian parliament 'Stortinget'; the latter are categorized simply as Norwegian.



Treebank Selection

Select a set of treebanks to work with. ?

Languages: **All** · Afrikaans (0/4) · Akkadian (0/4) · Akuntsu (0/1) · Albanian (0/1) · Amharic (0/3) · Ancient Greek (to 1453) (0/19) · Apurinã (0/1) · Arabic (0/16) · Armenian (0/3) · Assyrian Neo-Aramaic (0/2) · Bambara (0/3) · Basque (0/9) · Beja (0/1) · Belarusian (0/4) · Bhojpuri (0/2) · Breton (0/3) · Bulgarian (0/10) · Buriat (0/4) · Catalan (0/7) · Chinese (0/23) · Chukot (0/1) · Church Slavic (0/11) · Classical Armenian (0/1) · Coptic (0/5) · Croatian (0/9) · Czech (0/31) · Danish (0/11) · Dutch (0/16) · **English** (6/48) · Erzya (0/3) · Estonian (0/12) · Faroese (0/5) · Finnish (0/25) · French (0/33) · Galician (0/13) · **Georgian** (5/9) · **German** (6/30) · Gothic (0/9) · Guajajára (0/1) · Hebrew (0/9) · Hindi (0/12) · **Hungarian** (4/13) · Icelandic (0/6) · **Indonesian** (2/15) · Irish (0/10) · **Italian** (1/28) · Japanese (0/16) · K'iche' (0/1) · Kangri (0/1) · Karelian (0/2) · Kazakh (0/7) · Khunsari (0/1) · Komi (0/6) · Komi-Permyak (0/2) · Korean (0/10) · Latin (0/30) · Latvian (0/8) · Lithuanian (0/6) · Livvi (0/2) · Low German (0/1) · Makuráp (0/1) · Maltese (0/3) · Manx (0/1) · Marathi (0/4) · Mbyá Guaraní (0/4) · **Modern Greek (1453-)** (1/10) · Moksha (0/2) · Mundurukú (0/1) · Nayini (0/1) · Nigerian Pidgin (0/3) · Northern Kurdish (0/4) · Northern Sami (0/29) · **Norwegian** (5) · **Norwegian Bokmål** (47/58) · **Norwegian Nynorsk** (10/20) · Old English (ca. 450-1100) (0/5) · Old French (842-ca. 1400) (0/4) · Old Norse (0/8) · Old Russian (0/22) · Old Turkish (0/1) · Persian (0/10) · **Polish** (23/37) · **Portuguese** (1/25) · Romanian (0/15) · **Russian** (1/24) · Sanskrit (0/6) · Scottish Gaelic (0/2) · Serbian (0/4) · Skolt Sami (0/2) · Slovak (0/6) · Slovenian (0/16) · Sonha (0/1) · South Levantine Arabic (0/1) · Spanish (0/20) · Swedish (0/22) · Swedish Sign Language (0/5) · Swiss German (0/2) · Tagalog (0/4) · **Tamil** (1/10) · Telugu (0/4) · Thai (0/3) · Tupinambá (0/1) · **Turkish** (1/20) · Uighur (0/6) · Ukrainian (0/6) · Upper Sorbian (0/4) · **Urdu** (2/7) · Urubú-Kaapor (0/1) · Vietnamese (0/6) · Warlpiri (0/3) · Welsh (0/2) · Western Armenian (0/1) · Western Frisian (0/1) · **Wolof** (3/5) · Yoruba (0/3) · Yue Chinese (0/4) · Yupik (0/1)

Treebank Collections: **All** · **Acquis** (1/7) · Alpino (0/1) · BulTreeBank (0/1) · **CLARIN-PL** (5) · DELPH-IN (0/2) · GEGO (0/4) · **GeoGram** (4) · **HunGram** (4) · ISWOC (0/9) · JOS (0/1) · Menotec (0/8) · Mercurius (0/1) · **NAOB** (15) · **NDT** (2/4) · **NorGram** (58) · **NorGramBank** (40) · **POLFIE** (23) · PROIEL (0/10) · PaHC (0/2) · **ParGram** (11) · **ParTMA** (15) · Sami-open (0/15) · Sami-restricted (0/7) · **Sofie** (2/9) · **TIGER** (2/3) · TOROT (0/22) · Universal Dependencies 1.1 (0/19) · Universal Dependencies 1.2 (0/36) · Universal Dependencies 1.3 (0/53) · Universal Dependencies 1.4 (0/63) · Universal Dependencies 2.0 (0/63) · Universal Dependencies 2.1 (0/103) · Universal Dependencies 2.3 (0/130) · Universal Dependencies 2.5 (0/157) · Universal Dependencies 2.8 (0/200) · **WolGram** (3) · **XPar** (2)

Treebank Types: All · **lfg** (119) · constituency (19) · constituency-alpino (1) · dependency (49) · dependency-cg (864) · dependency-tuebadz (1) · hpsg (2)

Figure 3: The INESS user interface for treebank selection, with treebank type *lfg* chosen

A collection contains several treebanks with something in common, for instance that they were developed as part of a specific project, or that they consist of translations of the same text into different languages (including the source language text). A single treebank may belong to more than one collection. Type refers to the annotation type, such as LFG, HPSG, constituency, and dependency, and including subtypes of these. The user may click on any language, collection or type to make a first choice about which treebanks should be displayed.

In Figure 3 we see the effect of clicking on the type *lfg*; after this choice, only the languages and treebank collections that have LFG treebanks are displayed in boldface. Counting the boldfaced languages in Figure 3 shows that there are 16 languages that have LFG treebanks. After each language name, the numbers in parentheses indicate how many of the treebanks are LFG treebanks; for English, (6/48) means that six of 48 treebanks are LFG treebanks. In a similar manner,

under Treebank Collections, TIGER (2/3) means that two of the three treebanks in the collection called TIGER are LFG treebanks.

Once a first choice has been made by a user, a list of all treebanks matching that choice is displayed. When LFG is chosen, a total of 119 treebanks are listed. The top of this list is shown in Figure 4.[10] For each treebank, this overview shows its name, which collections it belongs to, its annotation type, its size (in sentences and words), whether it has been indexed for search, and the type of license (if any). The user may choose one or more treebanks by ticking off the boxes to the left of the treebank name; clicking on the name of one of the chosen treebanks brings the user to that treebank. When exploring a treebank for the first time, the user is asked to accept the license conditions.

Clicking on a treebank name brings the user to the *Sentence Overview* page for that treebank; the sentences are listed one per line together with information about their disambiguation status. Clicking on a sentence displays the *Sentence* page, where the analysis for that sentence is shown including the textual context the sentence occurs in (the previous and following three sentences).

Click on a treebank name below to proceed. All selected treebanks will be available for viewing and searching. | Show treebank descriptions

Selected (all | none)	Name	Collection	Type	Sentences	Words	Indexed		License
				17 828 129	252 023 248			
	English (eng)			**533**	**9 501**			
☐	**eng-jrc-acquis** (aligned)		lfg	94	2 188	yes		(Accepted)
☐	**eng-pargram** (aligned)	ParGram	lfg	101	658	yes	Ⓟ	CC-BY (Accepted)
☐	**eng-partma**	ParTMA	lfg	45	189	yes	Ⓟ	CC-BY (Accepted)
☐	**eng-partma-rat**	ParTMA	lfg	10	163	yes	Ⓟ	CC-BY (Accepted)
☐	**eng-partma-scorpion**	ParTMA	lfg	10	127	yes	Ⓟ	CC-BY (Accepted)
☐	**eng-partma-tempeval3**	ParTMA	lfg	273	6 176	yes	Ⓟ	CC-BY (Accepted)
	Georgian (kat)			**1 242**	**10 719**			
☐	**kat-mrs** (aligned)	GeoGram	lfg	106	374	no		(Accepted)
☐	**kat-pargram** (aligned)	GeoGram, ParGram	lfg	52	231	yes	Ⓟ	CC-BY (Accepted)
☐	**kat-partma**	ParTMA	lfg	34	106	yes	Ⓟ	CC-BY (Accepted)
☐	**kat-sofie** (aligned)	Sofie, GeoGram	lfg	1 025	9 915	yes		unspecified (Accepted)
☐	**kat-xpar** (aligned)	XPar, GeoGram	lfg	25	93	no		(Accepted)
	German (deu)			**20 278**	**255 589**			
☐	**deu-pargram** (aligned)	ParGram	lfg	102	644	yes	Ⓟ	CC-BY (Accepted)
☐	**deu-partma**	ParTMA	lfg	56	262	yes	Ⓟ	CC-BY (Accepted)
☐	**deu-partma-manifesto**	ParTMA	lfg	260	3 459	no	Ⓟ	CC-BY (Accepted)
☐	**deu-radio**		lfg	1 418	22 952	no		(Accepted)
☐	**deu-tiger**	TIGER	lfg	9 221	114 136	yes		(Accepted)
☐	**deu-tiger/subset**	TIGER	lfg	9 221	114 136	no		(Accepted)

Figure 4: Top of the list of treebanks after the type *lfg* has been chosen

[10] Treebank names in INESS begin with the three-letter ISO 639-3 code for the relevant language.

3.2 The TIGER treebank

The original TIGER treebank of German newspaper text (Brants et al. 2002, 2004) uses a hybrid annotation combining constituency and dependency information; part of it is also annotated with LFG structures. The constituency/dependency part of the treebank was constructed by two different methods. In one method a cascaded probabilistic parser was used in combination with manual annotation with the ANNOTATE tool (Brants & Plaehn 2000). The other method involved parsing with the German LFG grammar, followed by manual disambiguation; the XLE transfer system was employed to change the representations into the TIGER format (Zinsmeister et al. 2002). The LFG analyses were thus originally utilized in an experimental way to construct a more traditional treebank, but they now also constitute a useful resource as a standalone LFG treebank.

Figures 5 and 6 display the constituency/dependency and LFG analyses, respectively, for the sentence in (1). The URLs in parentheses in the captions are PIDs (persistent identifiers). They provide links to the analyses in the treebanks. Such links are persistent as long as the treebank they refer to remains available. For LFG treebanks, which are dynamic (they can be reparsed after changes are made to the grammar and/or lexicon), the PIDs are persistent in the sense that they provide a link to the current analysis of the same sentence, not necessarily exactly the same analysis shown here.

(1) German
Das Angebot ist bereits groß.
the offer is already large
'The offer is already large.'

The tree in Figure 5 contains information about both phrase structure and syntactic functions. The nodes in yellow boxes are phrasal categories, while the nodes in the blue boxes under the S node are syntactic functions: SB for subject, HD for head, MO for modifier and PD for predicate complement.

The c-structure in Figure 6 displays extensive unary branching—many nodes have only single daughters—and many complex category labels, i.e., c-structure nodes subscripted with features enclosed in square brackets. The latter device moves some of the feature complexity of the LFG grammar from the f-structure space into the context-free c-structure space, which improves parsing efficiency while maintaining the simplicity of the c-structure rules. In the f-structure we see that the SUBJ is also analyzed as the TOPIC, the predicate complement is analyzed as an XCOMP-PRED, and the modifier is analyzed as an ADJUNCT.

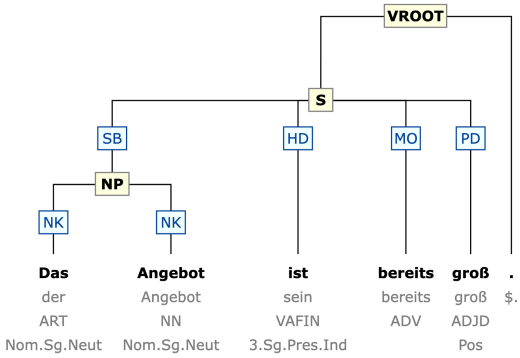


Figure 5: TIGER constituency/dependency analysis of (1) (http://hdl.handle.net/11495/D8B8-3970-851A-3@dep138682)

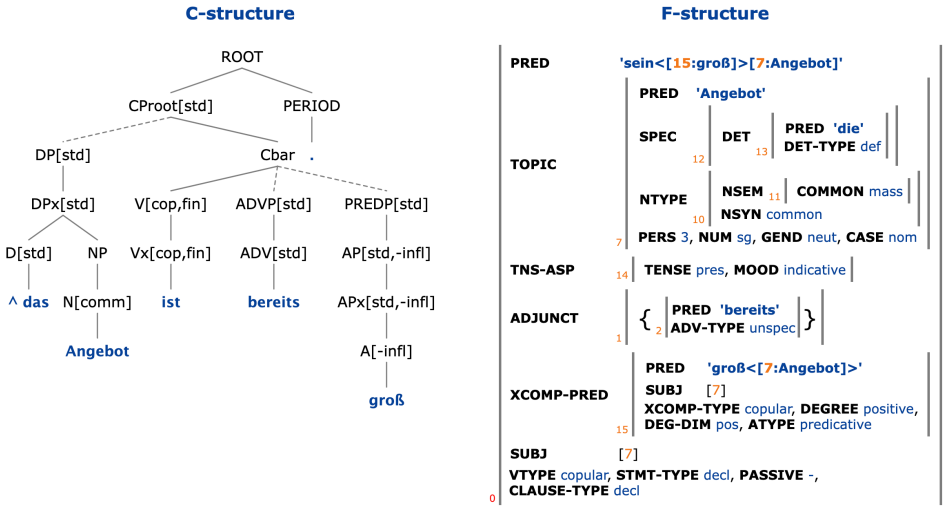


Figure 6: TIGER LFG analysis of (1) (http://hdl.handle.net/11495/D8B8-3970-851A-3@lfg41730)

3.3 The LFG Structure Bank for Polish

The LFG Structure Bank for Polish was built by parsing a corpus with the POLFIE grammar (Patejuk & Przepiórkowski 2012, 2014). This grammar was created by reusing context-free grammar rules written for another parser for Polish, Świgra, and adding annotations for building the f-structures. The corpus for the treebank is the one-million word subcorpus of the National Corpus of Polish[11] which has been manually annotated, the same subcorpus that was used for the previously annotated Składnica treebank.[12]

In INESS, the treebanks created by the POLFIE grammar are all in one large collection, also called POLFIE. This collection includes the LFG Structure Bank for Polish as well as other treebanks. The size of the POLFIE collection is 179,994 sentences and 2,022,026 words. Some of the subtreebanks in POLFIE are also in other collections: CLARIN-PL, ParGram and ParTMA.

Sample c- and f-structures from the POLFIE treebank for the sentence in 2 are given in Figure 7.

(2) Polish
Drzewo został-o ścięt-e wczoraj.
tree.NOM.SG.N get.3SG.N cut.NOM.SG.N yesterday
'The tree was cut down yesterday.'

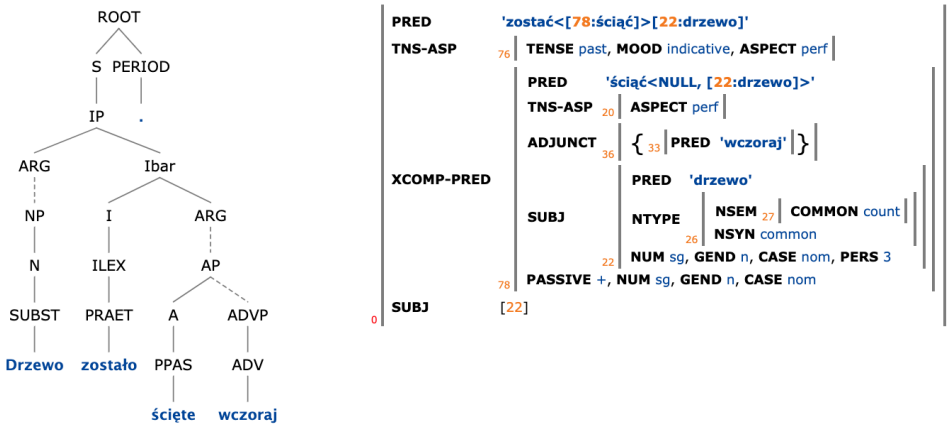


Figure 7: C- and f-structures for the Polish sentence in (2) (http://hdl.handle.net/11495/D8B8-3970-851A-3@lfg1411740)

[11] http://nkjp.pl/index.php?page=0&lang=1
[12] http://zil.ipipan.waw.pl/Składnica

In the c-structure we see some familiar categories such as A, ADV, ADVP, NP, N, I, Ibar, etc., but there are also categories which we might not immediately be able to identify, such as ILEX, PRAET and PPAS. Some terms in the f-structure may also be unfamiliar, such as NTYPE, NSEM and NSYN.[13] Treebank documentation should ideally be made available by treebank creators to assist users in exploring the treebank; unfortunately INESS lacks documentation for many treebanks.

An overview of all *indexed attributes* for each treebank may be found on the *Treebank Details* page. The indexed attributes are all labels used in the treebank annotation that can be searched for. For LFG treebanks, these attributes include *cat* (category) and *edge* (feature or attribute, in more standard LFG terminology). A screenshot of the top of this page is shown in Figure 8.

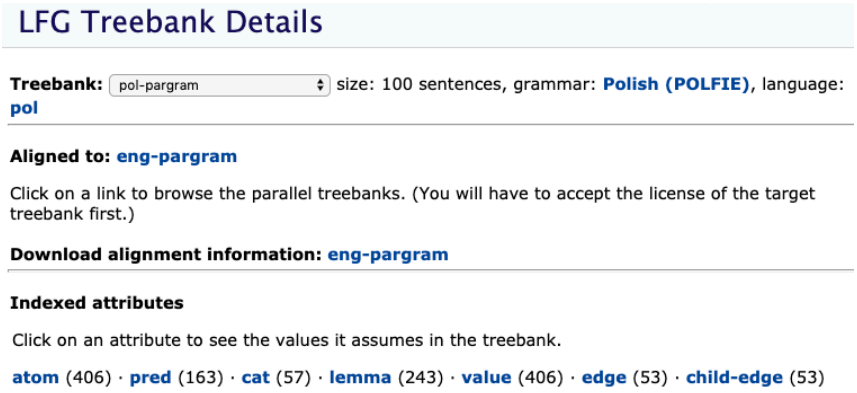


Figure 8: Treebank details for POLFIE

Clicking on *cat* and *edge* under *Indexed attributes* produces the lists in Figure 9. These lists are shown sorted according to frequency; we see that, for instance, the category NP occurs 236 times in this subcorpus (pol-pargram) consisting of 100 sentences.

3.4 NorGramBank: A Norwegian LFG parsebank

The INESS project had the twofold goal of building a treebanking infrastructure and of building the first large treebank for Norwegian. The result of the latter effort is the treebank collection NorGram, consisting of 15 million sentences (215

[13] These f-structure attributes also occur in Figure 6, and they illustrate the parallelism on the f-structure level achieved by the ParGram grammars; see Section 3.6.2.

cat, distinct values: 57

This table shows all values of the attribute, together with their corpus counts.

Sorted (•) by frequency ( ) alphabetically

236	NP	6	INF
201	N	5	MOD
199	SUBST	5	PPAS
156	IP	5	ADVP
115	ARG	5	RM
103	S	4	QUB
98	--	4	XPextr
86	ROOT	4	A-TYPE
83	PERIOD	4	PADJ
63	I	4	NUMP
63	ILEX	4	NUMbare
58	Ibar	4	NUM
56	AP	3	INT-MARK
56	PRAET	3	ADV
54	A	3	PRON
47	ADJ	3	PPRON3
40	AP-SURROUND-OR-NONE	2	SIEBIE
39	FIN	2	PACT
37	COMMA	2	MODPART
28	NEG	2	PADV
24	PP	2	XP_FOCUS
23	P	2	FM
23	PREP	2	CNEG
15	CP	2	CONJ
15	CPbare	1	IMPS
10	COMP	1	PCON
7	PRON-TYPE		
7	PSUBST		
7	IMPT		
7	XPsem		

edge, distinct values: 53

This table shows all values of the attribute, together with their corpus counts.

Sorted (•) by frequency ( ) alphabetically

587	-	12	TYPE
422	PRED	11	COMP
326	CHECK	10	_PREDICATIVE
326	_CAT	10	COMP-FORM
287	CASE	8	PFORM
285	NUM	8	OBL-STR
279	GEND	7	OBJ-TH
226	PERS	6	OBL
206	NTYPE	5	REFLEXIVE
206	NSYN	5	_RQR
206	NSEM	5	XCOMP
206	COMMON	4	OBL-LOCAT
132	SUBJ	4	ACM
117	TNS-ASP	2	OBL-COMPAR
117	ASPECT	2	UDF
103	TENSE	2	NEG-CONST
103	MOOD	2	COORD-FORM
98	_TOP	1	OBL-AG
75	$	1	XADJUNCT
72	OBJ	1	OBL-PERL
55	ADJUNCT	1	OBL-MOD
54	DEGREE	1	OBL-ADL
51	ATYPE		
36	PASSIVE		
28	NEG		
23	PTYPE		
20	POSS		
19	XCOMP-PRED		
14	_VOC		
13	CLAUSE-TYPE		

Figure 9: Values for the *cat* and *edge* attributes in POLFIE

million words) and by far the largest LFG treebank available in INESS. It was parsed with the eponymous grammar NorGram, a wide-coverage LFG grammar developed in the LOGON, TREPIL and INESS projects. Several versions of this grammar were constructed and used for parsebanking, including versions with c-structure pruning (Cahill et al. 2008). Some material was disambiguated manually with discriminants, but the bulk of the parsebank was disambiguated automatically through stochastic parseranking, based on the stored discriminants.

The collection NorGramBank (Dyvik et al. 2016) consists of a subset of the texts parsed in the NorGram collection. NorGramBank has more than 160 million words and consists of a variety of text types; while some newspaper texts were included, edited fiction and nonfiction texts were preferred because these have a higher language quality and fewer errors. Any error in a sentence, whether typographical, orthographical or grammatical, will result in a failure to find the intended analysis on parsing. Moreover, nonstandard constructions, interruptions, etc. may result in fragment analyses. Some NorGram texts were excluded from NorGramBank because the source texts had many OCR errors.

The text selection for the corpus was partially dependent on available resources. While published texts are valued sources for treebanks and other corpora, copyright restrictions must be taken into account. It is therefore paramount to clear permissions with rights holders before starting to work on texts. In the case of NorGram, several texts were obtained through the National Library of Norway. For some of these, copyright had expired. For newer texts, exceptional permission to use these with some restrictions was obtained from the government. Every corpus must be provided with metadata, including such information as provenance and conditions for use.

The Norwegian treebanks parsed with NorGram have proved useful for lexicography (see Section 4.4). Some NorGram treebanks have been specifically added for NAOB, a dictionary project by the Norwegian Academy for Language and Literature aimed at building a large dictionary for Norwegian Bokmål. In INESS, the collection called NAOB consists of 15 treebanks with a total of over 11 million sentences (161 million words).

The Norwegian example analyses shown in Figures 10, 11, 14, 15 and 18 are all from the NorGram treebanks.

3.5 Small treebanks for grammar development

Most of the small LFG treebanks in INESS are test suites used in various projects. GeoGram, HunGram and WolGram are collections of test suites used for the development of XLE grammars for Georgian (Meurer 2009), Hungarian (Laczkó et

al. 2013; Laczkó 2014) and Wolof (Dione 2014, 2019), respectively. Some of these test suites are parts of parallel treebanks (see 3.6). Other treebanks in these collections may only be available to their creators since they are work in progress and not at a stage where they may be useful to other researchers. Treebank developers decide whether they want to make their treebanks publicly available.

3.6 Parallel treebanks with LFG annotations

A parallel treebank is a collection of monolingual treebanks that are aligned with each other on the sentence level, and sometimes also on phrase and/or word levels. The most common type of parallel treebank involves one or more translations of a text that are aligned with the source text, but a parallel treebank can also have different annotations of the same text, for example a constituency annotation and a dependency annotation.

The user can select aligned parallel treebanks by choosing *Show only Parallel Treebanks* on the *Treebank Selection* page and selecting a collection from those that are then displayed in boldface. One of the treebanks to be examined is then chosen in the usual manner by clicking in the box next to the treebank name and subsequently clicking on the treebank name. From the *Sentence Overview* page, clicking on *Treebank Details* provides an overview of which other treebanks are aligned. Selecting one of those treebanks will start the display of parallel analyses for the two chosen languages.

The following subsections will present the XPAR Project (3.6.1), the treebanks developed in the Parallel Grammar Project (3.6.2), and other parallel treebanks containing LFG analyses (3.6.3).

3.6.1 The XPAR Project

Language Diversity and Parallel Grammars (XPAR) was a pilot project which aimed to determine to what extent the development of parallel deep grammars for typologically diverse languages may support the automatic derivation of high-quality parallel treebanks for those languages (Dyvik et al. 2009). Principles for phrase alignment and methodology for the automatic alignment of c-structures from manually aligned f-structures were developed in the project.

A small parallel test suite of translationally equivalent Georgian and Norwegian sentences was used in developing the alignment tool. An example of aligned sentences is provided in (3), and their sentence-aligned analyses are shown in Figure 10.

(3) a. Georgian
gia-s uqvars eka
Gia-DAT loves Eka.NOM
'Gia loves Eka.'

b. Norwegian
Jon elsker Maria.
Jon loves Maria
'Jon loves Maria.'

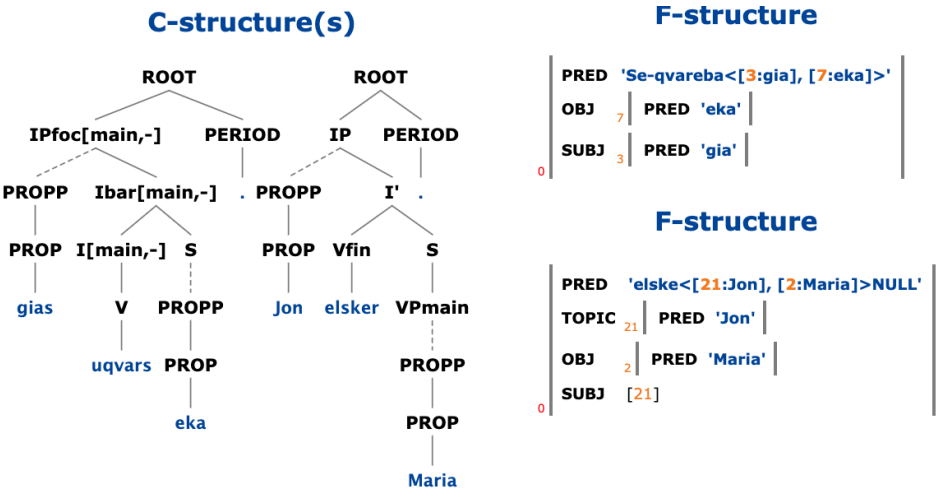


Figure 10: Sentence aligned c- and f-structures for the Georgian (http://hdl.handle.net/11495/D8B8-3970-851A-3@lfg51519) and Norwegian (http://hdl.handle.net/11495/D8B8-3970-851A-3@lfg60949) sentences in (3)

F-structures are manually aligned on the basis of translational correspondences at the level of predicate–argument structure. Subsidiary f-structures correspond if their predicates are in a translational relationship to one another. The alignment is done by dragging the index of one f-structure onto the corresponding index of the other f-structure. For instance, in Figure 10, the OBJ index 7 in the Georgian f-structure may be dragged onto the OBJ index 2 in the Norwegian one. This results in indices of the form $\boxed{n \rightarrow m}$, where n is the original index of that f-structure and m is the original index of the f-structure it is aligned with. Figure 11 shows the result of this manual alignment of f-structures, where the indices for the OBJ, SUBJ and main PRED have been aligned. Once the f-structures are aligned, the LFG Parsebanker automatically aligns the corresponding nodes in

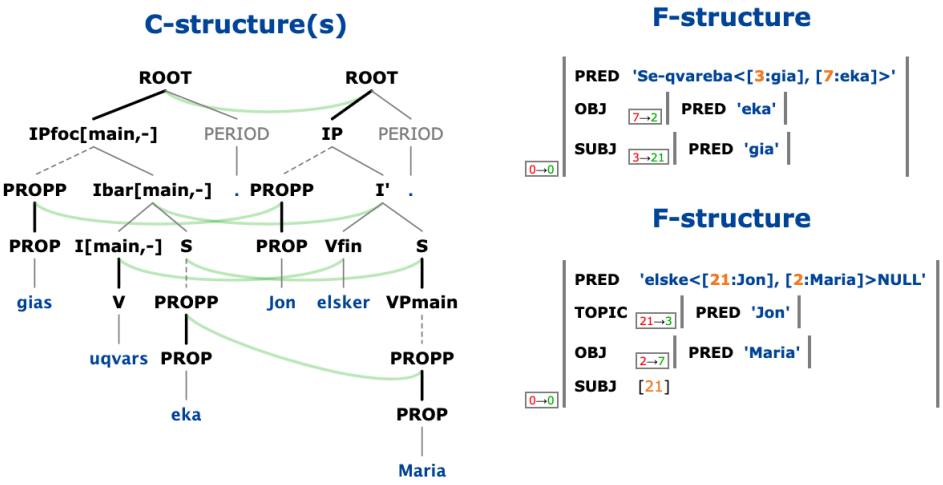


Figure 11: Word and phrase aligned c- and f-structures for the Georgian and Norwegian sentences in (3)

the c-structures, shown by the curved green lines. We see, for example, that the OBJ alignment in the f-structures results in the alignment of the PROPP nodes dominating *Eka* and *Maria* in the c-structures.

3.6.2 The Parallel Grammar Project treebanks

The Parallel Grammar Project (ParGram) is an international cooperative effort to develop parallel LFG grammars implemented in XLE (Butt et al. 1999, 2002). Originally three languages were involved in the project: English, French and German; later other languages joined, including Georgian, Hungarian, Indonesian, Japanese, Norwegian, Polish, Tamil, Turkish, Urdu and Wolof, among others. The main focus of the ParGram project was to develop and maintain linguistically motivated parallelism at the level of f-structure. Some of the ParGram participants have also been involved in the ParSem project, an effort to develop semantic structures based on the ParGram syntactic structures, with most of the ParSem systems using XLE's transfer system.

ParGram has created two parallel treebanks to support the aim of developing parallel LFG grammars. These treebanks consist of test suites encompassing various syntactic constructions. The English sentences were first agreed upon, and then translated into the other languages in the project. The first set of 50 sentences included such constructions as declaratives, interrogatives, imperatives, transitivity, passive, unaccusative, and subcategorized declaratives (Sulger et al.

2013). These sentences are included in the ParGram collection in INESS. Another set of sentences, concerned with tense, mode and aspect, constitutes the ParTMA collection.

(4) a. What did the farmer see?

b. German
Was sah der Bauer?
what saw the farmer
'What did the farmer see?'

Figure 12 shows word and phrase aligned c- and f-structures for the English and German sentences in (4). The f-structures for these sentences are practically identical, whereas the c-structures are quite different. This is both because the languages are different (English has *do*-support and German does not) and because the grammars for these languages have used quite different principles and techniques in writing the phrase structure rules. Still we see that most c-structure nodes are aligned. Since the XPAR principles align only translationally corresponding f-structures with PRED values, not all c-structure nodes can be aligned. The word *did* and the question marks only contribute features to the f-structure, not PRED values; these features are not shown here since the f-structures are displayed in PREDS only mode.

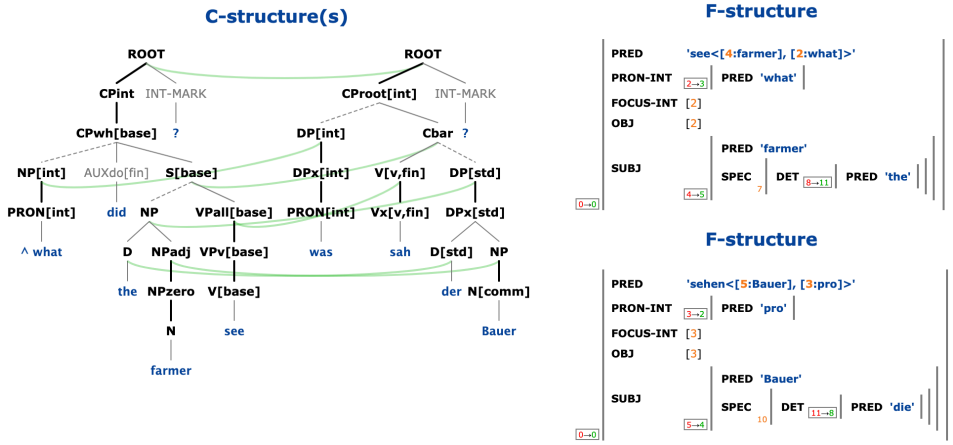


Figure 12: Word and phrase aligned c- and f-structures for the English (http://hdl.handle.net/11495/D8B8-3970-851A-3@lfg423651) and German (http://hdl.handle.net/11495/D8B8-3970-851A-3@lfg444239) sentences in (4)

3.6.3 Other parallel treebanks including LFG

Several projects have built parallel treebanks that include both LFG treebanks and treebanks of other types. Three such parallel treebanks are presented here.

The Sofie Parallel Treebank is a parallel corpus containing the first chapters of Jostein Gaarder's novel *Sofies verden* "Sophie's World". This text was chosen for treebanking because it is a well-written text that has been translated into a great number of languages. The Nordic Treebank Network[14] developed treebanks based on these texts for Danish, Estonian, German, Icelandic and Swedish in the period 2001–2005. The META-NORD project,[15] which ran from 2011 to 2013, had as one of its goals to promote the accessibility of treebanks, including some that had not been maintained and were no longer accessible (Losnegaard et al. 2013). An English treebank, originally developed in the SMULTRON project,[16] and a Georgian treebank, developed at Uni Computing in Bergen, Norway, were added to the Sofie collection. Two treebanks for Norwegian were also developed, one an LFG treebank and the other a constituency treebank with syntactic and functional categories. Only the Georgian and one of the Norwegian treebanks have LFG annotation; the rest of the treebanks have various types of constituency annotation. In the initial version of the LFG Sofie treebank for Norwegian, 73% of sentences received analyses. An in-depth study of the sentences that received full parses that were not entirely correct showed that 29% lacked the correct analysis because of grammar problems, while lexical problems accounted for 71%, with missing multiword expressions in the lexicon being the most important of these. Subsequent grammar and lexicon updates resulted in correct analyses for more than 90% of these sentences (Losnegaard et al. 2012).

The META-NORD Acquis Parallel Treebank is a small parallel corpus of translations of a European Union directive.[17] The EU languages Danish, Estonian, Finnish, Latvian and Swedish, as well as the non-EU languages Norwegian and Icelandic, have treebanks in the collection. All language pairs are aligned at sentence level. The Norwegian treebank contains LFG analyses, while the other languages have consistency or dependency annotations.

The Norwegian Dependency Treebank was developed by the National Library of Norway (Solberg et al. 2014); it is made available in INESS as the treebanks

[14] https://cl.lingfil.uu.se/~nivre/research/nt.html

[15] http://www.meta-net.eu/projects/meta-nord/

[16] https://www.cl.uzh.ch/en/texttechnologies/research/corpus-linguistics/paralleltreebanks/smultron.html

[17] Directive 2002/74/EC, from the Acquis Communautaire (AC), the total body of European Union law applicable in the member states.

named nob-ndt-dep (for Norwegian Bokmål) and nno-ndt-dep (for Norwegian Nynorsk). The treebank has also been converted to the Universal Dependencies (UD) annotation scheme (Øvrelid & Hohle 2016), creating the treebanks nob-ud-2.5-dep and nno-ud-2.5-dep. The same texts were parsed with NorGram to obtain LFG analyses, resulting in the treebanks nob-ndt-lfg and nno-ndt-lfg. The original dependency annotations were created automatically, but the analyses were then manually checked and corrected, resulting in a gold standard treebank. The dependency treebanks contain analyses for all sentences, while the LFG treebank has coverage for about 90% of the sentences. The analyses for the sentences that are covered in the LFG treebank are, however, much more detailed than those in the dependency treebanks. See 4.5 for more on UD treebanks, including a comparison with LFG analyses.

4 Exploring and exploiting LFG treebanks

4.1 INESS Search

Prior to the INESS project, there was no search tool that could perform search in LFG f-structures. INESS Search (Meurer 2012, 2020; Rosén et al. 2017) is a search tool that was developed in order to fill this need. It is a reimplementation and extension of TIGERSearch (Lezius 2002), a search system designed for the TIGER treebank (Zinsmeister et al. 2002; Brants et al. 2004). INESS Search retains the full functionality of TIGERSearch for querying constituency and dependency treebanks while extending its functionality in order to query fully general directed graphs like LFG f-structures; in addition, it can be used for search in HPSG treebanks. INESS Search supports almost full first-order predicate logic, including negation and existential and universal quantification, with the exception of universal quantification over disjunctions.

INESS Search is fully integrated in the INESS infrastructure and is used via its Web interface. There is extensive documentation for INESS Search online, both a walkthrough that describes how to get started searching in INESS treebanks,[18] and thorough documentation of the query language itself.[19]

In addition to extending TIGERSearch, INESS Search has implemented simplifications to the syntax of search expressions for more clarity. Suppose you want to find examples of NPs with AP modifiers that have embedded PPs, such as the German NP in (5). In TIGERSearch you could write the search expression in (6), whereas (7) is an equivalent abbreviated expression in INESS Search.

[18] https://clarino.uib.no/iness/page?page-id=INESS_Search_Walkthrough
[19] https://clarino.uib.no/iness/page?page-id=INESS_Search

(5) German
die von Slumbewohnern unerlaubt gebauten Lehmhütten
the by slum.dwellers illegally built mud.huts
'the mud huts illegally built by slum dwellers'

(6) `[cat="NP"] > #x:[cat="AP"] & #x > [cat="PP"]`

(7) `NP > AP > PP`

The TIGERSearch expression in (6) may be read as follows: "There is a node with the category NP that dominates a node #*x* with the category AP; this same AP node #*x* dominates a node with the category PP." Each node has a variable, but it does not always need to be expressed; in (6), it is necessary to specify through the use of an explicit variable that it is the same AP that is dominated by the NP and that dominates the PP, otherwise the search results would return all sentences where there is at least one NP dominating an AP and at least one AP dominating a PP. In the abbreviated INESS Search expression (7), this chaining is inferred, so that an explicit mention of the variable is not necessary in this case. Furthermore, as also shown in Table 1, node labels may be used directly in the search expression, lexical and terminal nodes need only be enclosed in double quotes, and atomic f-structure values only in single quotes. One of the search results for the search expression in (7) from the TIGER treebank, the NP in (5), is shown in Figure 13; the node labels mentioned in the search expression are highlighted in red in the graph.

Table 1: Some examples of abbreviated syntax in INESS Search

Expression	Abbreviation	Explanation
`[cat="NP"]`	`NP`	node labels
`[word="book"]`	`"book"`	lexical nodes in dependency treebanks; terminal nodes in LFG and phrase-structure treebanks
`[atom="sg"]`	`'sg'`	atomic f-structure values in LFG treebanks
`[PP > #x:NP & #x > PP]`	`PP > NP > PP`	chaining of relations

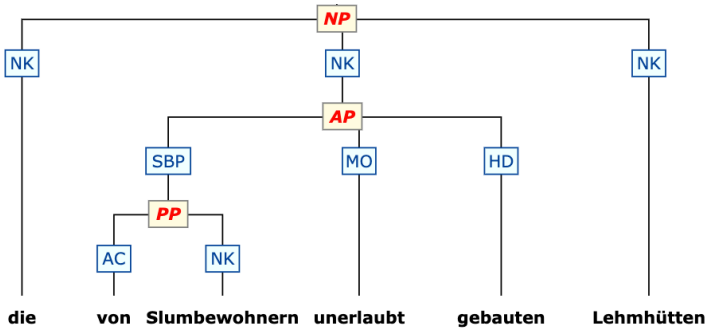


Figure 13: TIGER tree for (5) (http://hdl.handle.net/11495/D8B8-3970-851A-3@dep101299)

4.2 Querying with INESS Search

The formulation of well-targeted search expressions presupposes knowledge about the analyses in the treebank. One way of quickly gaining such knowledge is to use XLE-Web to parse sentences with the kind of grammatical phenomenon one is interested in and to study the analyses. Suppose that we want to search for passive sentences. The Norwegian passive sentence in (8) gets the analysis in Figure 14 when parsed in XLE-Web.

(8) Norwegian
Verden ble skapt av Gud.
world.DEF.SG was created by God
'The world was created by God.'

Examining the f-structure shows that the verb *skape* 'create' is the head of the XCOMP. It is a two-place predicate, with the PRED of the OBL-AG, *Gud* 'God', as its first argument, the agent. The XCOMP also has an attribute value pair 'PASSIVE +'. A simple search expression for passives with agent phrases can thus be formulated using these f-structure characteristics, as shown in (9).

(9) `#x >PASSIVE #y:'+' & #x >OBL-AG`

This expression may be read: "There is an f-structure #*x* which has an attribute PASSIVE with the value '+' (bound to #*y*), and this same f-structure #*x* also has an attribute OBL-AG."

The negation operator in INESS Search allows users to restrict searches with respect to properties that sentences should *not* have. The search expression in

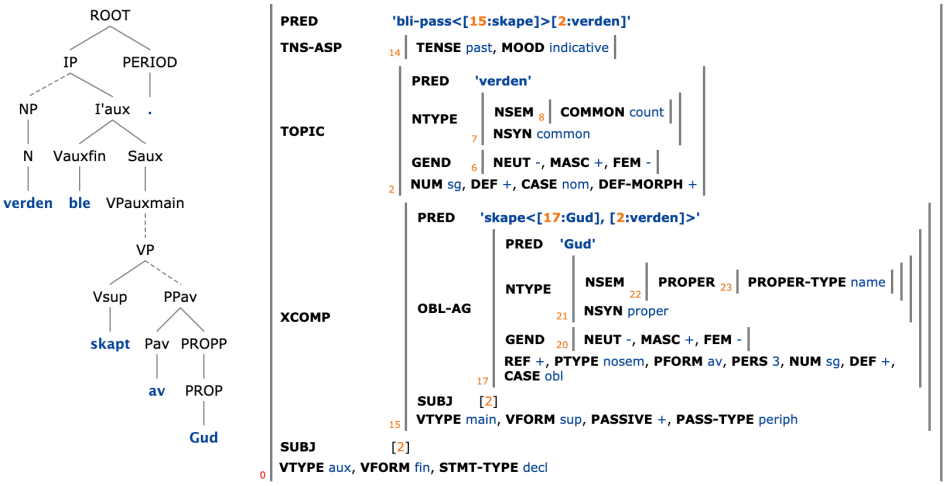


Figure 14: C- and f-structures for the passive sentence in (8) (http://hdl.handle.net/11495/D8B8-3970-851A-3@lfg6174124)

(10), where the exclamation point is the negation operator, searches for passives *without* agent phrases. The sentence in (11) is one of those found by this expression; its c- and f-structures are shown in Figure 15. The f-structure nodes that are named with explicit variables in the search expression are marked in red in the search result. In the f-structure we note that the XCOMP does not have an OBL-AG, and that the first argument of the main PRED is 'NULL'.

(10) `#x >PASSIVE #y:'+' & #x !>OBL-AG`

(11) Norwegian (http://hdl.handle.net/11495/D8B8-3970-851A-3@lfg599480)
Hvordan er verden skapt?
how is world.DEF.SG created
'How was the world created?'

4.3 An example-based introduction

For some researchers, INESS Search can be difficult to use, even with the simplifications that have been introduced. To assist users of NorGramBank in formulating search expressions, an example-based introduction to the search system has been written.[20] It is based on the Norwegian reference grammar *Norsk referanse-*

[20]This introduction, in Norwegian, is part of the INESS documentation: https://clarino.uib.no/iness/page?page-id=norgram-soek#innledning.

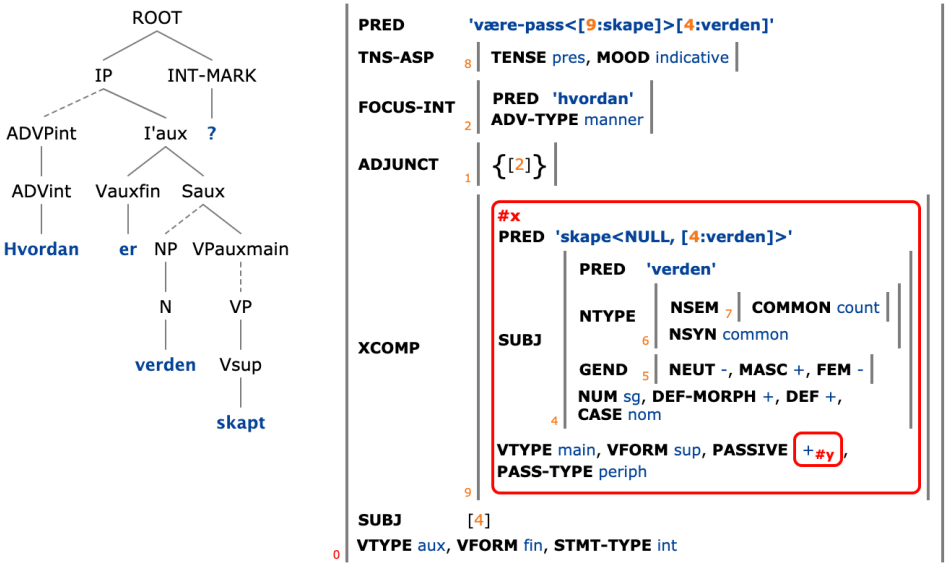


Figure 15: C- and f-structures for the passive sentence in (11)

grammatikk and the chapters and examples therein. Most researchers in Norwegian syntax will be familiar with the rather theory-neutral analyses in this book, and the goal is to provide them with LFG analyses of the constructions that are of interest, including the page numbers in the book where the constructions are treated. For each construction, the example-based documentation provides an LFG analysis of one sentence together with a commentary explaining the analysis. A search expression that will find the construction is provided, along with both a paraphrase and a lengthier prose explanation of the expression. Finally a list of a few matching sentences is presented.

A construction type that is difficult to search for without a treebank is relative clauses without complementizers. It would not be straightforward to find these in corpora which are not syntactically annotated, so this is a good illustration of the added value of treebanks. The search expression for relatives without complementizers is given in (12).

(12)
```
#x >(ADJUNCT $) #f >TOPIC-REL #g
& #f >OBJ #g & #f >CLAUSE-TYPE 'rel'
& !(#f >COMP-FORM)
& !(#x >PRON-TYPE 'free')
```

This search expression may be read: "An f-structure *#x* has an attribute ADJ-

JUNCT with a value that includes an f-structure #f; furthermore, #f has an attribute TOPIC-REL with the value #g, and an attribute OBJ with the same value #g; #f also has an attribute CLAUSE-TYPE with the value 'rel' and does not have an attribute COMP-FORM; the f-structure #x does not have an attribute PRON-TYPE with the value 'free' (the last specification ensures that free relatives will not be found)." An example sentence found by this expression in NorGramBank is given in (13), where the boldfaced relative clause *jeg så* lacks a complementizer.

(13) Norwegian
Alt **jeg så** var frontlykt-ene.
all I saw was headlight-DEF.PL
'All I saw was the headlights.'

4.4 Search with templates

A further simplification in INESS Search is the implementation of search templates, which abbreviate complete parameterized search expressions. For the Norwegian treebank NorGramBank, a number of such templates have been provided, primarily for the benefit of lexicographers.[21] Templates obviate the need for understanding an often complicated search expression, since users can choose one on the basis of a description of its intention, but they can examine the whole expression if desired. Templates are parameterized in the sense that the user can fill in values for one or more parameters, such as word or lemma forms, predicates, or grammatical features.

Suppose you want to find out how common nominal complement clauses with and without complementizers are after certain verbs. The template shown in Figure 16, named *AT-verbwithandwithout(@verb)*, may be used for this purpose. The user fills in the verb, in this case *fortelle* "tell, relate", and clicks on *Run query*. The results of the search are presented in a table, sorted according to whether they include the complementizer or not. We see that the vast majority of occurrences of complement clauses with this verb, 21,465 (97.5%), do have complementizers.

This can be compared with the results for the verb *tro* "think, be of the opinion", shown in Figure 17. For this verb the proportion of uses with the complementizer is only 33.8%. In this screenshot the user has clicked on the first row in the table, showing the number of occurrences for the verb without the complementizer (66,258). This brings up a list over all the sentences with this pattern. Here the user has clicked twice on *Next* in order to come to page 3; there are so many

[21]Documentation in Norwegian: https://folk.uib.no/hfohd/INESS-Sketch-veiledning-2020.pdf

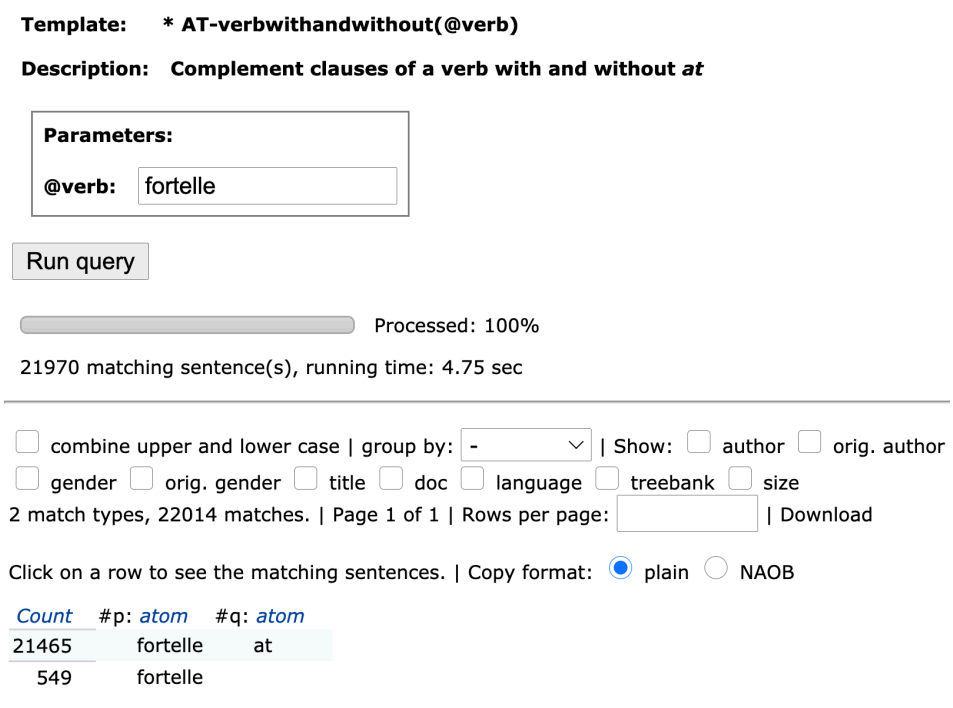
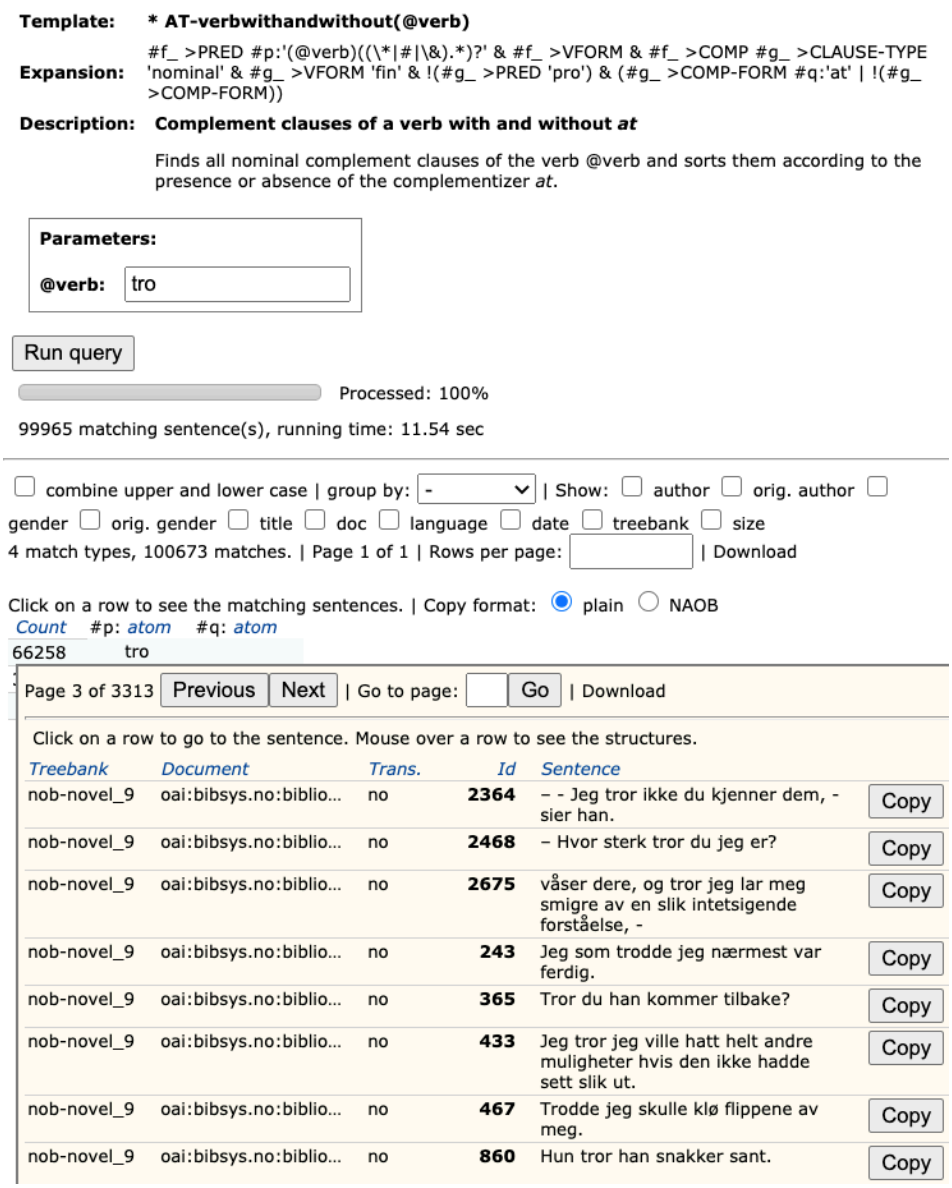


Figure 16: Template for nominal clause search with and without complementizer for the verb *fortelle* "tell, relate"

hits that the list consists of 3313 pages. When the user mouses over a sentence, a simplified f-structure is displayed to the right of the list. Clicking on a sentence brings the user to the *Sentence* page where the c-structure and the full f-structure are displayed. By default the quite complicated search expression which is used in this template is hidden, as in Figure 16. In Figure 17, the user has clicked on the template name, bringing up the expansion with the search expression. In Figure 17 a more detailed prose description is also displayed, obtained by clicking on the boldfaced, more compact, part of the description.

Rauset et al. (2021) provide concrete examples of the use of template search in NorGramBank for various dictionary projects in Norway. The lexicographers use templates to examine both the usage and frequency of words. The most common valency frames for verbs, as well as the most common prepositions and/or particles that they occur with, are examined by using the template *V-argframes(@V)*; this template also provides evidence about whether the verbs occur reflexively. The templates *ADJ-attrib-or-nominal(@ADJ)* and *V-attr-or-pred-ptc(@V)* provide evidence of the nominal and adjectival use of participles, which is sometimes the



Template: *** AT-verbwithandwithout(@verb)**

Expansion: #f_ >PRED #p:'(@verb)((*|#|\&).*)?' & #f_ >VFORM & #f_ >COMP #g_ >CLAUSE-TYPE 'nominal' & #g_ >VFORM 'fin' & !(#g_ >PRED 'pro') & (#g_ >COMP-FORM #q:'at' | !(#g_ >COMP-FORM))

Description: **Complement clauses of a verb with and without *at***

Finds all nominal complement clauses of the verb @verb and sorts them according to the presence or absence of the complementizer *at*.

Parameters:

@verb: tro

Run query

Processed: 100%

99965 matching sentence(s), running time: 11.54 sec

combine upper and lower case | group by: - | Show: author orig. author gender orig. gender title doc language date treebank size

4 match types, 100673 matches. | Page 1 of 1 | Rows per page: | Download

Click on a row to see the matching sentences. | Copy format: plain NAOB

Count	#p: atom	#q: atom
66258	tro	

Page 3 of 3313 Previous Next | Go to page: Go | Download

Click on a row to go to the sentence. Mouse over a row to see the structures.

Treebank	Document	Trans.	Id	Sentence	
nob-novel_9	oai:bibsys.no:biblio...	no	**2364**	– - Jeg tror ikke du kjenner dem, - sier han.	Copy
nob-novel_9	oai:bibsys.no:biblio...	no	**2468**	– Hvor sterk tror du jeg er?	Copy
nob-novel_9	oai:bibsys.no:biblio...	no	**2675**	våser dere, og tror jeg lar meg smigre av en slik intetsigende forståelse, -	Copy
nob-novel_9	oai:bibsys.no:biblio...	no	**243**	Jeg som trodde jeg nærmest var ferdig.	Copy
nob-novel_9	oai:bibsys.no:biblio...	no	**365**	Tror du han kommer tilbake?	Copy
nob-novel_9	oai:bibsys.no:biblio...	no	**433**	Jeg tror jeg ville hatt helt andre muligheter hvis den ikke hadde sett slik ut.	Copy
nob-novel_9	oai:bibsys.no:biblio...	no	**467**	Trodde jeg skulle klø flippene av meg.	Copy
nob-novel_9	oai:bibsys.no:biblio...	no	**860**	Hun tror han snakker sant.	Copy

Figure 17: Template for nominal clause search with and without complementizer for the verb *tro* "think, be of the opinion"

basis for the creation of separate entries for derived adjectives.

Targeted queries that provide evidence for colligations are useful when treating high-frequency words with many senses. The template *N-argofverbs(@N)* provides a list sorted by frequency of the verbs that occur with a certain noun as their first or second argument. Such results help lexicographers determine whether the sense distinctions made in older versions of the dictionaries are still reasonable, or whether there should be changes made by adding or removing distinctions, or for instance by promoting a sense that is now more common than previously.

An example of a word which was missing a sense is the reflexive verb *utmerke seg* 'distinguish oneself', which was defined as having only a positive connotation. The lexicographers, however, did not believe this to be accurate. The template *V-prepobj(@V,@P)* was used to examine which words occur as objects of the prepositions *med* 'with' and *ved* 'by'. The search results showed several occurrences of the noun *mangel* 'lack' as the object of *ved*; one of these examples is given in (14). This and similar searches provided empirical support for the establishment of a new subsense of the verb with a negative connotation.

(14) Norwegian (http://hdl.handle.net/11495/D8B8-3970-851A-3@lfg14979442)
Han vil ... utmerke seg med mangel på konsistens i sine
he will distinguish REFL with lack of consistency in his
handlingsvalg ...
action.choice
'He will ... distinguish himself with lack of consistency in his choice of actions ... '

4.5 Comparison of search in LFG and dependency treebanks

Dependency treebanks are the most widely used type of treebanks, notably through the Universal Dependencies (UD) initiative.[22] The UD treebanks are grounded in dependency grammar, which assigns dependency relations between words, and does not analyze phrases and constituency relations (Tesnière 1959). An important early dependency treebank was the Prague Dependency Treebank (Hajič et al. 2001). Among the treebanks provided by INESS, dependency treebanks are the most numerous (250), with the UD treebanks accounting for most of these (200). The latest version in INESS at the time of writing is 2.8. INESS also keeps earlier versions, making it possible to track progress between versions.

[22] https://universaldependencies.org

The LFG and UD analyses of the sentence in (15) are shown in Figures 18 and 19. For both treebanks, information about lemma, part of speech and morphological features may be displayed (by clicking on the word for the dependency treebank, or by clicking on the preterminal node for the LFG treebank). The c-structure in Figure 18 shows the hierarchical phrase structure of the sentence, labeled with a rich inventory of syntactic categories. The corresponding f-structure encodes syntactic functions, grammatical features, and predicate–argument relations, as represented in the semantic forms of the verbs. The dependency structure in Figure 19 is shallower and less detailed than the LFG structure. Dependencies between words are shown by labeled arrows that go from a word to its dependents.

(15) Norwegian
Han hadde aldri vært lykkeligere.
he had never been happier
'He had never been happier.'

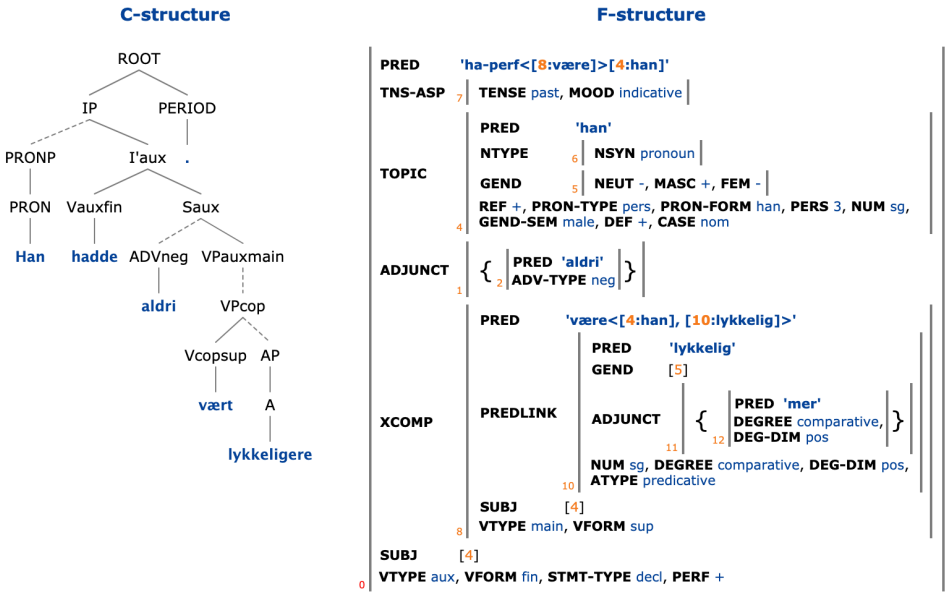


Figure 18: LFG analysis of the sentence in (15) (http://hdl.handle.net/11495/D8B8-3970-851A-3@lfg4292653)

The deeper analysis in an LFG treebank improves the search possibilities as compared with a dependency treebank. Rosén et al. (2020) compares search in the UD version of the Norwegian Dependency Treebank with the same texts in

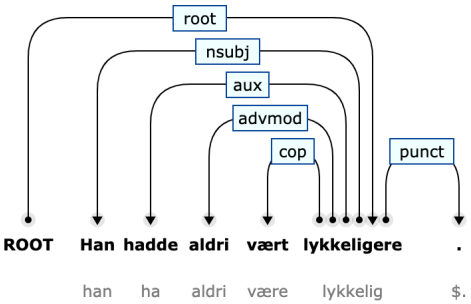


Figure 19: UD analysis of the sentence in (15) (http://hdl.handle.net/11495/D8B8-3970-851A-3@dep8965528)

NorGramBank. The example given there is searching for the first argument of verbs. This may be done straightforwardly in an LFG treebank, but it is much more difficult in a dependency treebank since predicate–argument structure is not encoded there. The first argument of a verb can be the subject of an active verb or of a predicative present participle, the agent phrase of a passive verb, or the head of an attributive present participle. And since the UD guidelines allow for several ways of annotating some of these possibilities, creating a search expression to capture them is extremely complicated. For more detail on this comparison, see Rosén et al. (2020).

5 Conversion between LFG treebanks and other treebanks

Besides pure parsebanking with a grammar, other approaches have been used to construct treebanks by converting between formalisms or by enriching treebanks with additional information. The Universal Dependencies initiative is in some ways similar to ParGram in that both approaches aim at assigning common annotations to comparable items and structures across languages.

Since dependency relations may be labeled as grammatical functions such as subject and object, dependency structures have a resemblance to f-structures in LFG. The PARC 700 Dependency Bank is a treebank in dependency format based on the English LFG grammar developed at PARC (King et al. 2003). The corpus was created only to make a dependency bank. LFG analyses were transformed to dependency graphs, but no LFG treebank per se was created.

The TIGER corpus, mentioned in 3.2, utilized the large-scale German LFG grammar of the ParGram project for the semiautomatic creation of TIGER treebank annotations. The grammar was used for full parsing, followed by semi-

automatic disambiguation and automatic transfer into the treebank format (Zinsmeister et al. 2002). The hybrid representation structure of TIGER, combining constituent analysis and functional dependencies, benefited from information in the c-structures and f-structures provided by the LFG grammar.

Conversely, an LFG treebank may be created by enriching phrase-structure oriented treebank resources with functional structures, as suggested by Frank et al. (2003) and Cahill (2004). For more on grammar induction, see Cahill & Way forthcoming [this volume].

Forst (2003) describes a method for converting the TIGER treebank to a test-suite for the German LFG ParGram grammar. The conversion utilizes the machine translation transfer system in XLE.

Recently, detailed algorithms for the conversion from LFG analyses to dependency structures were proposed by Meurer (2017) and Przepiórkowski & Patejuk (2020). While the latter follow the more standard assumption that f-structures provide a good basis for developing dependency trees, the former takes c-structures as the starting point, but combines this with information from f-structures.

6 Conclusion

This chapter has provided an introduction to LFG treebanks, illustrated throughout with the tools and visualizations of the INESS treebanking infrastructure. The process of developing an LFG grammar in tandem with a treebank through incremental parsebanking has been described. Both large and small LFG parsebanks for a number of languages have been presented. Several different methods for searching LFG treebanks with INESS Search have been explained: users can write search expressions themselves with the aid of XLE-Web and the INESS Search documentation; they can find search expressions for the phenomena they are interested in by consulting the example-based search documentation; and they can use search templates that only require filling in one or more search items. LFG treebanks have been compared with other treebanks, and it has been shown that the more detailed and sophisticated annotation in LFG treebanks provides richer opportunities for research than simpler annotations.

While INESS has already been developed over more than a decade, the system, and especially its interface, will continue to evolve. Consequently, future interactions may be slightly different from the interactions and screen displays shown in this chapter.

Although LFG treebanks are certainly valuable resources for research and development, building an LFG treebank is a time consuming and expensive undertaking, especially for a language for which no large-coverage LFG grammar and

lexicon yet exist. However, the task is made somewhat easier with the help of the LFG Parsebanker as described above, and INESS is open to making more treebanks accessible for research and development.

Acknowledgments

I thank Koenraad De Smedt, Stefanie Dipper, Helge Dyvik, Jonas Kuhn, Paul Meurer, Agnieszka Patejuk, Heike Zinsmeister and three anonymous reviewers for helpful comments and suggestions. Thanks are also due to Mary Dalrymple for all of her help and her boundless patience.

References

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith & Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation* 2(4). 597–620. DOI: 10.1007/s11168-004-7431-3.

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.

Brants, Thorsten & Oliver Plaehn. 2000. Interactive corpus annotation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC'00)*. http://www.lrec-conf.org/proceedings/lrec2000/pdf/334.pdf.

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.

Cahill, Aoife. 2004. *Parsing with automatically acquired, wide-coverage, robust, probabilistic LFG approximations*. Dublin: School of Computing, Dublin City University. (Doctoral dissertation).

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 & Andy Way. Forthcoming. Treebank-driven Parsing, Translation and Grammar Induction using LFG. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

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.

Dione, Cheikh M. Bamba. 2014. LFG parse disambiguation for Wolof. *Journal of Language Modelling* 2(1). DOI: 10.15398/jlm.v2i1.81.

Dione, Cheikh M. Bamba. 2019. Clause structure, pro-drop and control in Wolof: an LFG/XLE perspective. *Nordic Journal of African Studies* 28(3). 1–26.

Dyvik, Helge, Paul Meurer, Victoria Rosén & Koenraad De Smedt. 2009. Linguistically motivated parallel parsebanks. In Marco Passarotti, Adam Przepiórkowski, Sabine Raynaud & Frank Van Eynde (eds.), *Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories*, 71–82. Milan: EDUCatt.

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.

Forst, Martin. 2003. Treebank conversion – Establishing a testsuite for a broad-coverage LFG from the TIGER treebank. In *Proceedings of the 4th international workshop on linguistically interpreted corpora (LINC-03) at EACL 2003*, 205–216. Association for Computational Linguistics. https://www.aclweb.org/anthology/W03-2404.

Forst, Martin & Tracy Holloway King. Forthcoming. Computational implementations and applications. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

Frank, Anette, Louisa Sadler, Josef van Genabith & Andy Way. 2003. From treebank resources to LFG f-structures. In Anne Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*, 367–389. Dordrecht: Kluwer Academic Publishers. DOI: 10.1007/978-94-010-0201-1_21.

Hajič, Jan, Barbora Vidová-Hladká & Petr Pajas. 2001. The Prague Dependency Treebank: Annotation structure and support. In *Proceedings of the IRCS Workshop on Linguistic Databases*, 105–114. Philadelphia.

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 Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*. Budapest: Association for Computational Linguistics.

Laczkó, Tibor. 2014. 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, György Rákosi, Ágoston Tóth & Gábor Csernyi. 2013. Nyelvtanfejlesztés, implementálás és korpuszépítés: A HunGram 2.0 és a HG-1 Treebank legfontosabb jellemzői [Grammar development, implementation and corpus construction: Key features of HunGram 2.0 and the HG-1 Treebank]. In Attila Tanács & Veronika Vincze (eds.), *IX. Magyar Számítógépes Nyelvészeti Konferencia konferenciakötete*, 85–96. Szeged: JATEPress.

Lezius, Wolfgang. 2002. TIGERSearch – Ein Suchwerkzeug für Baumbanken. In Stephan Busemann (ed.), *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*. Saarbrücken.

Losnegaard, Gyri Smørdal, Gunn Inger Lyse, Anje Müller Gjesdal, Koenraad De Smedt, Paul Meurer & Victoria Rosén. 2013. Linking Northern European Infrastructures for improving the accessibility and documentation of complex resources. In Koenraad De Smedt, Lars Borin, Krister Lindén, Bente Maegaard, Eiríkur Rögnvaldsson & Kadri Vider (eds.), *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013, May 22–24, 2013, Oslo, Norway. NEALT Proceedings Series 20* (Linköping Electronic Conference Proceedings 89), 44–59. Linköping University Electronic Press.

Losnegaard, Gyri Smørdal, Gunn Inger Lyse, Martha Thunes, Victoria Rosén, Koenraad De Smedt, Helge Dyvik & Paul Meurer. 2012. What we have learned from Sofie: Extending lexical and grammatical coverage in an LFG parsebank. In Jan Hajič, Koenraad De Smedt, Marko Tadić & António Branco (eds.), *Proceedings of the META-research workshop on advanced treebanking at LREC'12*, 69–76. Istanbul. http://www.lrec-conf.org/proceedings/lrec2012/workshops/12.LREC%202012%20Advanced%20Treebanking%20Proceedings.pdf.

Maxwell, John T., III & Ronald M. Kaplan. 1993. The interface between phrasal and functional constraints. *Computational Linguistics* 19. 571–590.

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.

Meurer, Paul. 2012. INESS-Search: A search system for LFG (and other) treebanks. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 404–421. Stanford, CA: CSLI Publications.

Meurer, Paul. 2017. From LFG structures to dependency relations. *Bergen Language and Linguistics Studies* 8. DOI: 10.15845/bells.v8i1.1341.

Meurer, Paul. 2020. Designing efficient algorithms for querying large corpora. *Oslo Studies in Language* 11(2). 283–302. DOI: 10.5617/osla.8504.

Meurer, Paul, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Gunn Inger Lyse, Gyri Smørdal Losnegaard & Martha Thunes. 2013. The INESS treebanking infrastructure. In Stephan Oepen, Kristin Hagen & Janne Bondi Johannessen (eds.), *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)* (Linköping Electronic Conference Proceedings), 453–458. Linköping University Electronic Press. http://www.ep.liu.se/ecp/085/043/ecp1385043.pdf.

Meurer, Paul, Victoria Rosén & Koenraad De Smedt. 2020. Interactive visualizations in INESS. In Miriam Butt, Annette Hautli-Janisz & Verena Lyding (eds.), *LingVis: Visual analytics for linguistics*, 55–85. Stanford, CA: CSLI Publications / University of Chicago Press.

Øvrelid, Lilja & Petter Hohle. 2016. Universal dependencies for Norwegian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA).

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 & 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.

Rauset, Margunn, Gyri Smørdal Losnegaard, Helge Dyvik, Paul Meurer, Rune Kyrkjebø & Koenraad De Smedt. 2021. Words, words! Resources and tools for lexicography at the CLARINO Bergen Centre. Manuscript under review.

Rosén, Victoria. Forthcoming. LFG treebanks. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

Rosén, Victoria & Koenraad De Smedt. 2022. Managing treebank data with the infrastructure for the exploration of syntax and semantics (INESS). In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.), *MIT open handbook of linguistic data management* (Open Handbooks in Linguistics), 499–512. Cambridge, MA: The MIT Press.

Rosén, Victoria, Koenraad De Smedt & Paul Meurer. 2006. Towards a toolkit linking treebanking to grammar development. In *Proceedings of the 5th Workshop on Treebanks and Linguistic Theories*, 55–66.

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, Helge Dyvik, Paul Meurer & Koenraad De Smedt. 2020. Creating and exploring LFG treebanks. In Miriam Butt & Ida Toivonen (eds.), *Proceedings of the LFG '20 conference*, 328–348. Stanford, CA: CSLI Publications.

Rosén, Victoria, Helge J. Jakhelln Dyvik, Paul Meurer & Koenraad De Smedt. 2017. Exploring treebanks with INESS search. *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*. Linköping Electronic Conference Proceedings 29(131). 326–329. http://www.ep.liu.se/ecp/131/048/ecp17131048.pdf.

Rosén, Victoria, Paul Meurer & Koenraad De Smedt. 2005. Constructing a parsed corpus with a large LFG grammar. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '05 conference*. Stanford, CA: CSLI Publications.

Rosén, Victoria, Paul Meurer & Koenraad De Smedt. 2007. Designing and implementing discriminants for LFG grammars. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '07 conference*, 397–417. Stanford, CA: CSLI Publications.

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.

Solberg, Per Erik, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen & Janne Bondi Johannessen. 2014. The Norwegian Dependency Treebank. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

Sulger, Sebastian, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh M. Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Agnieszka Patejuk, Özlem Çetinoglu, I Wayan Arka & Meladel Mistica. 2013. ParGramBank: The ParGram Parallel Treebank. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics*, vol. 1, 550–560. Sofia: Association for Computational Linguistics. https://www.aclweb.org/anthology/P13-1054.pdf.

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.

Zinsmeister, Heike, Jonas Kuhn & Stefanie Dipper. 2002. TIGER transfer — Utilizing LFG parses for treebank annotation. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '02 conference*, 427–447. Stanford, CA: CSLI Publications.

Chapter 3

LFG and Australian languages

Rachel Nordlinger
University of Melbourne

Australian languages exhibit many interesting grammatical properties and have featured in LFG-related research since the earliest days of the framework. In this chapter I survey the features of Australian languages that have featured most prominently in work within LFG, and show how they argue strongly for the parallel architecture of LFG and in particular the separation of functional relations at f-structure from phrasal constituency and linearity at c-structure. These morphosyntactic features include nonconfigurationality and flexible word order, the role of morphology in encoding grammatical relations, case stacking, valence-changing phenomena and complex predicates. I show how the flexibility afforded by LFG's parallel architecture, which separates c-structure from f-structure with a many-to-many mapping between them, allows for a natural and explanatory account of these properties of Australian languages. In return, the empirical questions prompted by these theoretical analyses and their predictions have led to a more detailed understanding of the intricate grammatical structures of various Australian languages, and explain the appeal of the LFG formalism for fieldworkers engaged in Australian language documentation.

1 The languages of Australia

Across the continent of Australia there are hundreds of Indigenous languages. The literature typically cites upwards of 800 named language varieties, which can be grouped into 250-300 distinct languages (Koch & Nordlinger 2014b), but it is not always straightforward to determine language differences from dialectal differences and so these numbers are approximate to a certain extent.[1] Prior to

[1]It is important to note that these >800 language varieties are considered different languages by Indigenous communities themselves, and thus the grouping of these into a smaller number of 'distinct languages' is a purely linguistic enterprise.

Rachel Nordlinger. Forthcoming. LFG and Australian languages. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press. 

the English invasion of Australia, these languages were spoken across a population of perhaps 750,000 to one million people, which highlights the enormous linguistic diversity of Indigenous Australia. In many cases languages were maintained by very small populations (e.g. 40–50 people), and the largest populations speaking a single language variety were probably no bigger than 4000 people. Linguistic diversity is highly valued culturally for its indexical relationship to heritage, identity and group membership (Evans 2007) and is not an impediment to communication, since high degrees of multilingualism were (and often still are) the norm across Indigenous Australia, with individuals typically speaking up to 4–6 languages of the surrounding area, as well as understanding others, given widespread practices of receptive multilingualism (Singer 2018).

Australian languages are generally considered by linguists to all be related to one another, although the detailed comparative work needed to establish this is still underway. Such research is confounded by a number of factors, the most significant of which is the extraordinary time depth (perhaps as much as 65,000 years) that Indigenous people have been living on and moving around the continent, with almost no written records of any of the languages prior to the last 200 years or so, and few detailed descriptions until substantially later. Research to date has established that the Australian languages can be grouped into around 25 different language families. One of these, the *Pama-Nyungan* family, covers approximately 85 percent of the continent, stretching from the south-west of Western Australia all the way to the tip of Cape York in far north Queensland. The other families, known collectively as the *non-Pama-Nyungan* families, are concentrated in the northern parts of Western Australia and the top half of the Northern Territory, but higher order groupings amongst these non-Pama-Nyungan families have not yet been clearly established.

The sociolinguistic situation varies enormously across these hundreds of languages and their communities (DITRC et al. 2020). Some languages remain strong, and are used by their communities as the daily language of communication and learned as first languages by the children. Many others are used fluently only by older members of the community, with younger generations having passive and varying degrees of partial knowledge of the language; while many other languages, particularly those from the areas most heavily populated by non-Aboriginal populations since the 19th century, have no first language speakers at all and are instead in the process of being relearned and revived by community members from (often scant) historical materials.

Australian languages are relatively similar phonologically (Fletcher & Butcher 2014) but exhibit greater variation in grammatical organisation. While all Australian languages are morphologically complex, we can see them as falling into

two broad grammatical types which we can loosely call dependent-marking and head-marking (Nichols 1986) (although most of the head-marking languages have some dependent-marking as well, and some of the dependent-marking languages have bound pronominal clitics cross-referencing verbal arguments). The Pama-Nyungan languages are dependent-marking languages with grammatical relations primarily encoded through case marking. These languages are generally morphologically ergative languages, and have elaborate case systems that cover a range of grammatical and semantic case functions. Examples such as the following are typical.

(1) Jiwarli
Ngatha tharla-laartu ngurru-martu-nha pirru-ngku.
1SG.ERG feed-USIT old.man-GROUP-ACC meat-ERG
'I used to feed the old men with meat.' (Austin 2001: 310)

(2) Jiwarli
Wuru ngunha tharrpa-rninyja ngarti-ngka kajalpu-la...
stick.ACC that.ACC insert-PST inside-LOC emu-LOC
'(He) inserted the stick inside the emu…' (Austin 2001: 315)

However, some other Pama-Nyungan languages combine a robust case-marking system with bound pronominal clitics cross-referencing verbal arguments, as illustrated in the following examples:

(3) Bilinarra
Liward-ba=nggu=lu garra nyununy gajirri-lu.
wait-EP=2MIN.OBJ=3AUG.SBJ be.PRS 2MIN.DAT woman-ERG
'The women are waiting for you.' (Meakins & Nordlinger 2014: 121)

(4) Bilinarra
Jamana-lu=rni=warla=rna=rla ma-ni warlagu=ma nyila=ma,
foot-ERG=ONLY=FOC=1MIN.SBJ=3OBL do-PST dog(ACC)=TOP that(ACC)=TOP
garndi-murlung-gulu.
stick-PRIV-ERG
'I kicked the dog of his with just my foot, not with a stick.' (Meakins & Nordlinger 2014: 121)

The head-marking languages largely belong to non-Pama-Nyungan families of northern Australia and encode core grammatical relations primarily through verbal morphology. Some of these are characterised as polysynthetic since verbs

can be so morphologically complex that they can stand alone as a single complex clause, and may even allow noun incorporation as in (5). The polysynthetic, head-marking languages of Australia have minimal grammatical case marking, although many still employ case for semantic case functions. Polysynthetic Australian languages include Bininj Gun-wok (Evans 2003) and Murrinhpatha (Blythe 2009; Nordlinger 2017; Mansfield 2019), as illustrated in the following examples.

(5) Bininj Gun-wok
Nga-ban-marne-yawoih-dulk-djobge-ng.
1SG.SBJ-3PL.OBJ-BEN-again-tree-cut-PST.PFV
'I cut the tree/wood for them again', or 'I cut another tree for them.'
(Evans & Sasse 2002: 2)

(6) Murrinhpatha
Puddan-wunku-rlarl-deyida-ngime=pumpanka.
3DU.SBJ.SHOVE.NFUT-3DU.OBJ-drop-in.turn-PC.F=3DU.SBJ.GO.NFUT
'They (dual sibling) are dropping them (paucal, female, non-sibling) off, one after the other, as they go along.' (Blythe 2009: 134)

Australian languages exhibit many interesting grammatical properties that have been the focus of much theoretical and typological discussion, including flexible word order, syntactic and morphological ergativity, elaborate case systems and case marking, nominal classification, complex verb structures, polysynthesis, noun incorporation, grammaticalised expression of kin relations, and many more – see the overviews and discussions in Dixon (2002), Koch & Nordlinger (2014a), and Bowern (2021) for more details. It is not possible for me to do justice to all of this work here, so in this chapter I focus on the features of Australian languages that have featured most prominently in work within the LFG framework.

2 Overview of work on Australian languages in LFG

Australian languages have featured in LFG-related research since the early days, beginning with Jane Simpson's PhD work on Warlpiri (Simpson 1983). The non-configurational clausal structure of languages like Warlpiri, first discussed by Hale (1981, 1982, 1983), argues strongly for the parallel architecture of LFG and in particular the separation of functional relations at f-structure from phrasal constituency and linearity at c-structure. Languages like Warlpiri provide clear support for the idea that the same f-structure information can be realised across

different languages with wildly diverse c-structures. This is illustrated by comparing Figure 1 and Figure 2 (based on Bresnan et al. 2016: 3–4), where we see that the same f-structure can correspond to both the highly configurational c-structure of English, and the flat non-configurational c-structure of Warlpiri. Warlpiri in addition allows multiple alternative word orders in c-structure, all of which correspond to this same f-structure.[2]

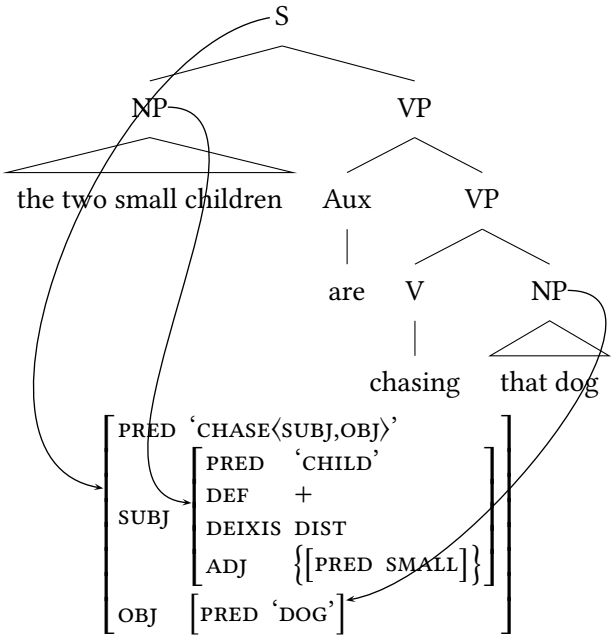


Figure 1: Simple c-structure/f-structure correspondences in English

While early work in LFG focussed on Warlpiri (Simpson 1983; Simpson & Bresnan 1983; Simpson 1991) subsequent work has brought in empirical data from a number of other Australian languages including Jiwarli (Austin & Bresnan 1996), Wambaya (Nordlinger & Bresnan 1996; Nordlinger 1998b), Dyirbal (Manning 1996), Wagiman (Wilson 1999), Kayardild (Evans & Nordlinger 2004), Wubuy (Baker & Nordlinger 2008; Baker et al. 2010), Anindilyakwa (van Egmond 2008), Arrernte (Dras et al. 2012) and Murrinhpatha (Seiss & Nordlinger 2010; Seiss 2013). The morphosyntactic properties of Australian languages that have

[2]Any order of words and categories in the c-structure given in Figure 2 is grammatical and semantically equivalent, as long as *ka=pala* remains second position. See (8) below for further exemplification.

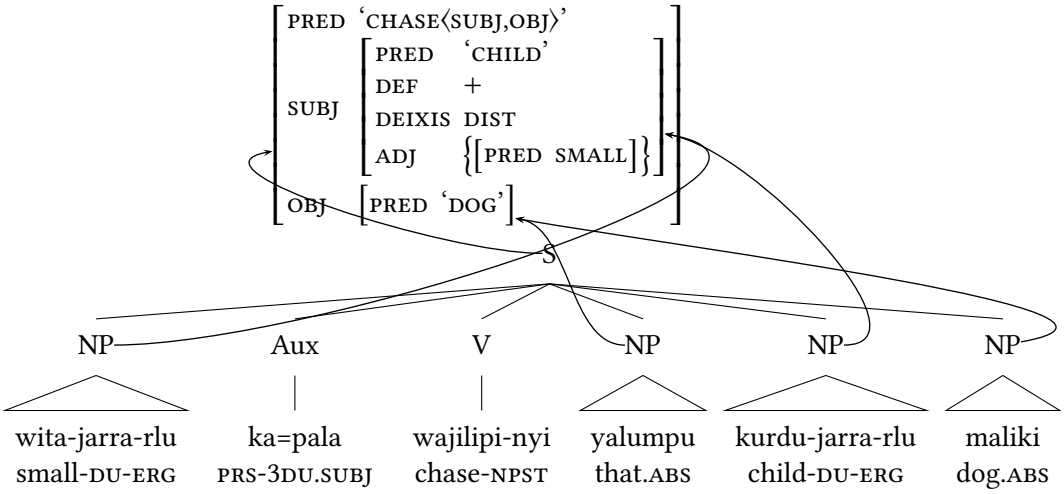


Figure 2: Simple c-structure/f-structure correspondences in Warlpiri

been discussed and analysed in this LFG literature range from clause structure and especially non-configurationality (Simpson 1991; Austin & Bresnan 1996; Nordlinger & Bresnan 2011; Snijders 2015) (see also Andrews forthcoming(a) [this volume]); the role of morphology in encoding grammatical relations (Nordlinger & Bresnan 2011; Nordlinger 1998b) including pronominal incorporation and verbal agreement (Austin & Bresnan 1996) and case marking (Simpson 1991; Andrews 1996; Nordlinger 1998b; Andrews 2017); and flexible noun phrase structure and discontinuity (Simpson 1991; Sadler & Nordlinger 2006a, 2010; Snijders 2016) to other morphosyntactic interactions such as the marking of tense/aspect/mood on NPs (Nordlinger & Sadler 2004a); valency-changing phenomena (Austin 1997; Seiss & Nordlinger 2010) and complex predicates (Wilson 1999; Andrews & Manning 1999). These will all be discussed in more detail in Section 3.

Given the morphological complexity of Australian languages – some head-marking and even polysynthetic, and others heavily dependent-marking – the LFG work on Australian languages has focussed largely on the morphology-syntax interface. It is here that the data from Australian languages contributes most to the development of LFG theory, and where the flexibility afforded by LFG's parallel architecture, which separates c-structure from f-structure with a many-to-many mapping between them, allows for a natural and explanatory account of the morphosyntax of Australian languages. Crucial to this flexibility is the fact that words (and therefore morphology) can contribute information directly to the f-structure alongside, or instead of, f-structure information coming

from the c-structure. This enables the framework to capture the cross-linguistic generalisation that languages rich in morphological structure, such as the Australian languages, often make less use of phrase structure – a generalization that Bresnan (2001: 7) captures with the slogan "morphology competes with syntax" – essentially words and phrases are different means of encoding the same grammatical relations (Nordlinger & Bresnan 2011). The unification-based architecture of LFG allows for compatible information from different structural sources to integrate into a single f-structure. The independence of grammatical functions from c-structure, along with features such as economy of expression (allowing for the optionality of c-structure heads) and an exocentric S category have contributed to the analysis of Australian languages in the framework, as discussed in more detail in Section 3. In return, the empirical questions prompted by these theoretical analyses and their predictions have led to a more detailed understanding of the intricate grammatical structures of various Australian languages, and explain the appeal of the LFG formalism for fieldworkers engaged in Australian language documentation.

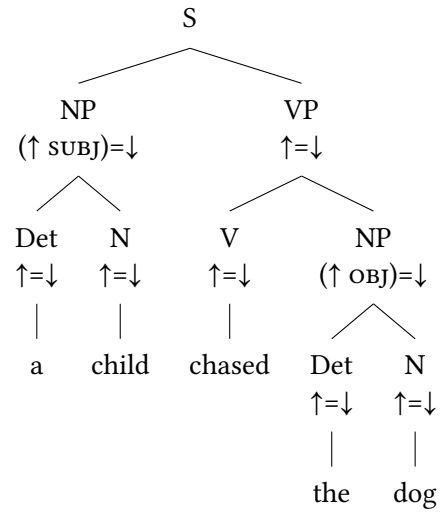
3 Phenomena analysed within LFG

3.1 Non-configurational clausal structure

Simpson (1983) observes that "Warlpiri, a Pama-Nyungan language spoken in Central Australia, is a language in which the burden of representing the relations between predicates and arguments ... is borne by the morphology rather than the syntax." (p. 18). Thus, many properties commonly associated with constituent structure in languages such as English are instead associated with morphological structure in Warlpiri, including the encoding of grammatical relations such as subject and object. In a configurational language like English grammatical relations can be associated with positions in a hierarchical constituent structure, as shown in (7b).

(7) a. A child chased the dog.

b.



In a language such as Warlpiri, on the other hand, constituent structure plays no role in identifying the grammatical relations of subject and object, as shown by the fact that the NPs in the Warlpiri sentence in (8) can appear in any position in the clause without affecting the meaning. Rather, it is the case marking, the morphological information carried by the nominals themselves, that plays the role of encoding grammatical relations information. In (8), the presence of the ergative case on 'child' and absolutive case on 'dog' unambiguously identifies the former as the subject NP and the latter as the object NP, irrespective of their positions in the constituent structure.

(8) Warlpiri
Kurdu-ngku maliki wajilipu-ngu.
child-ERG dog.ABS chase-PST
'A child chased the dog.' (Mary Laughren, pers. comm.)

Maliki wajilipu-ngu kurdu-ngku
Wajilipu-ngu kurdu-ngku maliki
Maliki kurdu-ngku wajilipu-ngu
Kurdu-ngku wajilipu-ngu maliki
Wajilipu-ngu maliki kurdu-ngu.

The disassociation of grammatical functions from hierarchical constituent structure in this way is known as 'non-configurationality', and discussion of Warlpiri, as well as some other dependent-marking Australian languages such as Wambaya

(Nordlinger 1998b) and Jiwarli (Austin & Bresnan 1996; Austin 2001) has been central to debates about the ways in which such languages are syntactically distinct from more configurational languages, and how best to represent these differences in formal syntactic theory. Hale (1983) identifies three key properties of Warlpiri syntax that he considers to be characteristic of its non-configurational structure: 'free word order' as illustrated in (8), 'the use of syntactically discontinuous expressions', whereby elements relating to the same grammatical relation can be discontinuous in the clause (9), and 'extensive use of null anaphora', which allows for the free omission of argument NPs (10).

(9) Warlpiri
Wawirri kapi=rna panti-rni yalumpu
kangaroo.ABS AUX=1.SG.SUBJ spear-NPST that.ABS
'I will spear that kangaroo.' (Hale 1983: 6)

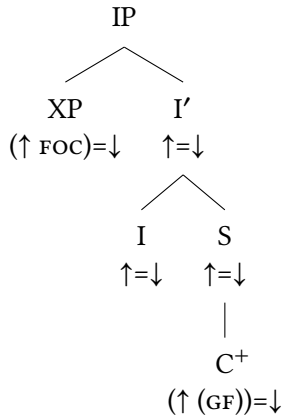
(10) Warlpiri
Panti-rni ka.
spear-NPST AUX
'He/she is spearing him/her/it.' (Hale 1983: 7)

Each of these properties illustrates the fact that grammatical relations in Warlpiri (and other similarly non-configurational languages) are not uniquely determined by the phrase structure position of the relevant argument NP. The fact that argument NPs can grammatically appear in any position in the clause, and that there can be multiple, discontiguous positions associated with the same grammatical function suggest that standard endocentric principles of X′ Theory do not apply uniformly in these languages. The free omission of argument NPs indicates that information about grammatical relations can be encoded elsewhere in the clause (e.g. as part of the verb's lexical and/or morphological content), not necessarily by phrase structure position. Austin & Bresnan (1996) show that these three properties vary independently of each other and that a language may be non-configurational without allowing 'discontinuous NPs', for example; rather, what is definitional for non-configurationality is the fact that grammatical relations are not directly defined by phrase structure position.

Simpson (1983, 1991) (also Hale 1983; Austin & Bresnan 1996; Nordlinger 1998b) argue that such non-configurationality supports a theoretical model in which phrase structure constituency is separated from functional relations, as in LFG (Austin & Bresnan (1996) call this the 'dual structure' hypothesis). The principles of c-structure in LFG, in addition to the standard categories determined by X′

theory, include a non-projective category S, distinguished from these other categories by the fact that it is not headed by something of the same category as itself (exocentric) (Bresnan 2001) (see also Andrews forthcoming(a) [this volume]). The availability of this category in c-structure allows for languages to have non-hierarchical, non-configurational phrase structures. Since this category is non-projective and exocentric, it can have a head of any category and, since it is not subject to the constraints of X′ Theory, it can dominate multiple constituents not bearing the typical relations of sisters in endocentric structures. Thus, S may define a totally flat phrase structure in which all constituents are sisters – all daughters of the clause – and functional annotations are assigned freely to all constituents, thereby capturing properties such as free word order and the possibility of discontinuous constituents. Following the analysis of Warlpiri c-structure provided by Austin & Bresnan (1996), the c-structure of a basic Warlpiri sentence can be given as in (11):

(11)



Where C = X^0, or NP

In this structure a non-configurational category S is generated as a sister to I within IP.[3] I is the position of the auxiliary, and the (optional) specifier of IP carries the discourse function of FOCUS.[4] The annotation (↑ (GF)) = ↓ associated with the constituents of S indicates that the functional annotations ↑ = ↓ (the head relation) and (↑ GF) = ↓ (where GF stands for the disjunction of all possible grammatical functions) are assigned freely within S (Simpson 1991; Austin & Bresnan 1996). Effectively this means that no specific functions are assigned within

[3]In some non-configurational languages such as Jiwarli (Austin & Bresnan 1996) there may be no evidence for an IP so that the top node of a clause is simply S.

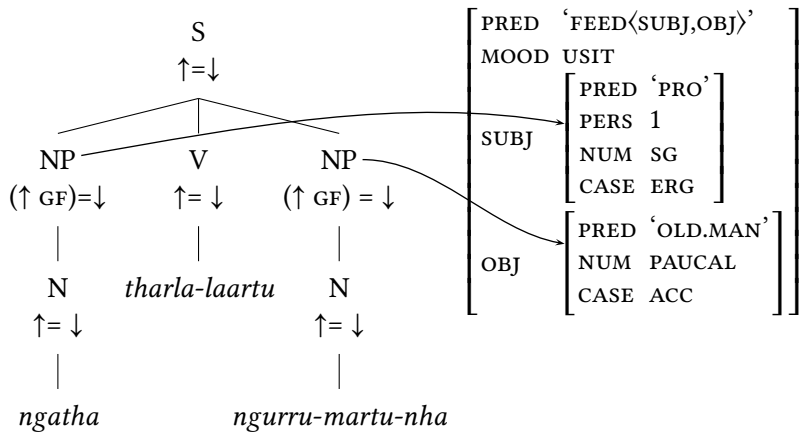
[4]For more detailed discussion of the motivation for this structure, and especially the second position of the auxiliary see Austin & Bresnan (1996).

S at all. Rather, it is the information encoded in the morphology in conjunction with the principles of Completeness and Coherence (see Belyaev forthcoming [this volume]) that ensures a grammatical c-structure and f-structure.

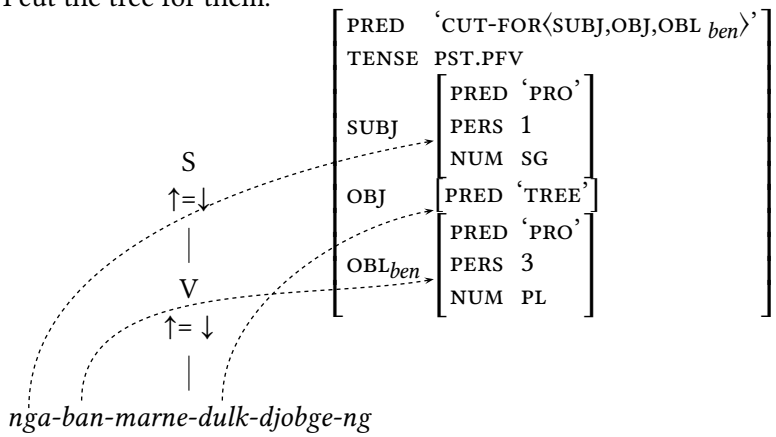
The principle of Economy of Expression in LFG (Bresnan 2001) states that all phrase structure nodes are optional unless they are required by independent principles. This allows for the possibility of null anaphora, since argument NPs are not required if the relevant grammatical function information is also contributed by morphological information (or by something else in the structure). Grammatical relations such as SUBJECT and OBJECT are encoded at f-structure and, since words in LFG can contribute information to the f-structure in the same way as syntactic phrases (Belyaev forthcoming [this volume]), words can contribute grammatical function information to f-structure directly, without the need for such information to also be reflected in the phrase structure. This provides a great deal of flexibility in terms of where and how different languages may encode grammatical function information, and even allows for languages to express it redundantly in both the phrasal syntax and the morphology, as long as the information is compatible under unification at f-structure (see Nordlinger (1998b: Chapter 3) for detailed discussion). Dependent-marking non-configurational languages such as Jiwarli (Austin 2001) encode grammatical function information primarily in case marking morphology, while head-marking non-configurational languages such as Bininj Gun-wok do this through verbal morphology. Warlpiri, with both case marking and pronominal argument clitics, combines both of these options. These options are illustrated in the following examples.

(12) Jiwarli
Ngatha tharla-laartu ngurru-martu-nha.
1SG.ERG feed-USIT old.man-GROUP-ACC
'I used to feed the old men.' [5]

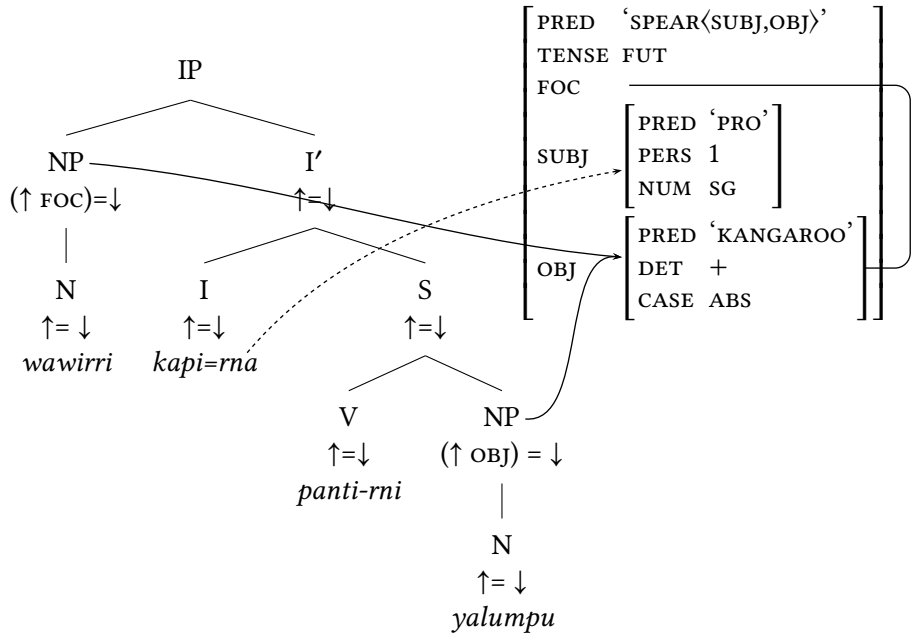
[5]This example is modified from (Austin 2001: 310). I have left the adjunct phrase *pirru-ngku* 'with meat' out here just to simplify the structures for presentational purposes.



(13) Bininj Gun-wok
Nga-ban-marne-dulk-djobge-ng.
1SG.SBJ-3PL.OBJ-BEN-tree-cut-PST.PFV
'I cut the tree for them.'



(14) Warlpiri
Wawirri kapi=rna panti-rni yalumpu
kangaroo.ABS AUX=1.SG.SUBJ spear-NPST that.ABS
'I will spear that kangaroo.'



In head-marking languages, grammatical function information is encoded as part of the inflected verb's lexical entry, associated with verbal agreement morphology in the usual way (see Haug forthcoming [this volume], also Börjars et al. (2019: Chapter 4) for detailed exemplification). Consider a Bininj Gun-wok verb such as that given in (15), the lexical entry for which is shown in (16). Following Bresnan & Mchombo (1987), the PRED values associated with the verbal morphology are optional to capture the fact that the verb can combine optionally with external argument NPs. When there are no co-referential NPs in the clause, the principle of Completeness will ensure that the PRED 'PRO' features are present, since otherwise the resulting f-structure will be incomplete, containing a SUBJECT and OBJECT lacking PRED features. In the presence of a co-referential NP, however, as in example (15), the OBJ PRED feature will be omitted since it will not be able to unify with the PRED value of the external object NP (see Belyaev forthcoming [this volume] for discussion of the Uniqueness principle and PRED values). This flexibility captures the fact that such verbal morphology can function as pronominal arguments, and also as agreement morphology in the presence of external NPs (see Toivonen forthcoming [this volume]).

(15) Bininj Gun-wok

Abanmani-na-ng bininj.
1SG.SUBJ:3DU.OBJ-see-PST.PFV man
'I saw the two men.' (Evans 2003: 417)

(16) *abanmaninang* (↑ PRED) = 'SEE⟨SUBJ,OBJ⟩'
(↑ TENSE) = PST.PFV
((↑ SUBJ PRED) = 'PRO')
(↑ SUBJ PERS) = 1
(↑ SUBJ NUM) = SG
((↑ OBJ PRED) = 'PRO')
(↑ OBJ PERS) = 3
(↑ OBJ NUM) = DU

In dependent-marking languages, such as Jiwarli and Warlpiri, grammatical function information is encoded by case morphology. There have been a number of different approaches to capturing this in LFG. Simpson (1983, 1991) assumes a verb-mediated approach, where verbs select for the case values of their arguments in their lexical entries. Thus, a verb such as *panti-* 'spear' would include in its lexical entry (↑ SUBJ CASE)=ERG and (↑ OBJ CASE)=ABS, which then must unify with the case value of the NP in the f-structure, constrained by the principles of Completeness and Coherence. Nordlinger & Bresnan (2011) supplement the verb-mediated approach with case conditionals of the type in (17), thus capturing the generalisation that there is a direct relationship between case and the encoding of grammatical functions.

(17) (↓ CASE) = κ ⇒ (↑ GF) = ↓

The idea is that each case value (represented here by κ) is associated in the grammar with a set of grammatical functions. For example, the case conditional for the Warlpiri ergative case might look as in (18), which specifies that an element with ergative case is to be associated with the subject grammatical function:

(18) (↓ CASE) = ERG ⇒ (↑ SUBJ) = ↓

Thus, by virtue of its case value each NP is assigned a grammatical function (or set of possible functions). In addition, verbs and other lexical predicators select for the case features of their arguments.[6] The unification of the possible functions of the NP and the requirements of the predicator, in conjunction with the general

[6]In the majority of cases this is predictable from the argument structure of the verb, so can be covered by a lexical rule.

principles of Uniqueness, Completeness and Coherence, ensures that the NPs in the c-structure are associated with the appropriate grammatical functions in the corresponding f-structure.

For example, a transitive verb stem such as *wajilipi-* 'chase' requires that its subject have ergative case and its object have absolutive case, thus corresponding to an f-structure such as the following:

(19)
$$\begin{bmatrix} \textsc{pred} & \text{'}\textsc{chase}\langle\textsc{subj},\textsc{obj}\rangle\text{'} \\ \textsc{subj} & \begin{bmatrix}\textsc{case} & \textsc{erg}\end{bmatrix} \\ \textsc{obj} & \begin{bmatrix}\textsc{case} & \textsc{abs}\end{bmatrix} \end{bmatrix}$$

The only f-structures for a sentence headed by this verb stem that satisfy Completeness and Coherence will be those in which an absolutive NP is identified with the OBJ grammatical function and an ergative NP is identified with the SUBJ grammatical function. Thus, the f-structure for the sentence in (20a) is that given in (20b).

(20) a. Warlpiri
Kurdu-ngku maliki wajilipu-ngu.
child-ERG dog.ABS chase-PST
'A child chased the dog.'

b.
$$\begin{bmatrix} \textsc{pred} & \text{'}\textsc{chase}\langle\textsc{subj},\textsc{obj}\rangle\text{'} \\ \textsc{tense} & \textsc{pst} \\ \textsc{subj} & \begin{bmatrix}\textsc{pred} & \text{'}\textsc{child}\text{'} \\ \textsc{case} & \textsc{erg}\end{bmatrix} \\ \textsc{obj} & \begin{bmatrix}\textsc{pred} & \text{'}\textsc{dog}\text{'} \\ \textsc{case} & \textsc{abs}\end{bmatrix} \end{bmatrix}$$

Nordlinger (1998b) provides a third approach to the analysis of case morphology and its role in encoding grammatical relations, known as 'constructive case'. This is discussed in more detail in Section 3.3.

The discussion of non-configurationality in Australian languages and its treatment in LFG has been expanded in more recent years to integrate information structure and its interaction with different word order possibilities. Simpson (2007) focusses on Warlpiri and the pragmatic constraints on its different word orders; this is also discussed for Jiwarli in Austin (2001). Snijders (2015) builds on and expands the earlier LFG work to provide a typology of configurationality that integrates information structure into the analysis, and extends the discussion beyond just the languages of Australia.

3.2 Flexible NP structure

Another feature common to many Australian languages that has been the subject of theoretical work in LFG is flexibility of NP structure.[7] While some researchers (including Hale 1983) consider this phenomenon to be central to the issue of non-configurationality in fact, as Austin & Bresnan (1996), Nordlinger (1998b) and others have argued, the two phenomena are logically distinct, although they may co-exist in a single language of course, as found in Warlpiri (Hale 1983), Jiwarli (Austin 2001), Wambaya (Nordlinger 1998b) and many other Australian languages. It is possible, however, for a language to be non-configurational at the clausal level while having strictly defined and non-flexible NPs. This is what we find in the Australian languages Kayardild (Evans 1995) and Murrinhpatha (Mujkic 2013), for example, both of which have clearly defined NP constituents with little or no discontinuity, while allowing great word order freedom at the clausal level and no clear association of grammatical relations with phrase structure. Languages such as these are thus non-configurational as discussed in Section 3.1 despite not allowing discontiguous nominal phrases.

The flexibility of NP structure in (some) Australian languages has been addressed within the LFG literature with regards to two different aspects. The first of these is NP discontinuity, the general LFG approach to which was discussed in Section 3.1 above (see also Snijders 2016 and Börjars & Lowe forthcoming [this volume]). The second is nominal juxtaposition – whereby many semantically different NP structures, including coordination, are expressed through the simple juxtaposition of nominals in seemingly flat NP structures (Sadler & Nordlinger 2006a, 2010). Sadler & Nordlinger (2010) provide the following illustrative examples:

(21) Coordination (Nyangumarta)
Pala-nga ngatu jarri-nya-pinti-ngi, mima-nikinyi-yi puluku,
that-LOC stationary INCH-NMLZ-ASSOC-LOC wait.for-IPFV-3PL.SBJ 3DU.DAT
kujarra kangkuru-jirri waraja yalapara.
two kangaroo-DU one goanna.
'And there, on the finishing line, the two kangaroos and one goanna waited for those two.' (Sharp 2004: 315)

(22) Generic-Specific (Yidiny)

[7] Recent work investigating this aspect of Australian languages in more detail includes Louagie & Verstraete (2016), Louagie (2020) and Reinöhl (2020). The details of these typological studies have not yet been fully addressed within LFG analyses.

Gana mayi jimirr jula:lin.
TRY vegetable(ABS) yam(ABS) dig.GOING.IMP
'Go and try to dig some yams up!' (Dixon 1977: 247)[8]

(23) Apposition (Wambaya)
Garidi-ni bungmanyi-ni gin-amany yanybi.
husband-ERG old.man-ERG 3SG.M.SBJ-PST.TWD get
'(Her) old man husband came and got (her).' (Nordlinger 1998a: 133)

(24) Inclusory (Kayardild)
Nga-rr-a kajakaja warra-ja thaa-th.
1-DU-NOM daddy.NOM go-ACT return-ACT
'Daddy and I will go' (lit. 'We two, including daddy, will go') (Evans 1995: 249)

Sadler & Nordlinger (2010) draw on the standard LFG treatment of coordination (Dalrymple & Kaplan 2000) to account for asyndetic coordination structures such as (21). Thus, the coordination structure is licensed by the c-structure rule in (25), where X is a metavariable ranging over N and NP, and the syntactic resolution of PERS and NUM features that is characteristic of coordination is captured by the template @NP-CNJT associated with each coordinand, which is defined as in (26). The resulting f-structure of the coordinated NP in (21) is given in (27).

(25) $\mathrm{X} \longrightarrow \underset{\substack{\downarrow \in \uparrow \\ @\text{NP-CNJT}}}{\mathrm{X}} , \underset{\substack{\downarrow \in \uparrow \\ @\text{NP-CNJT}}}{\mathrm{X}}$

(26) NP-CNJT: $(\downarrow \text{IND PERS}) \subseteq (\uparrow \text{IND PERS})$
$(\downarrow \text{IND NUM}) \subseteq (\uparrow \text{IND NUM})$

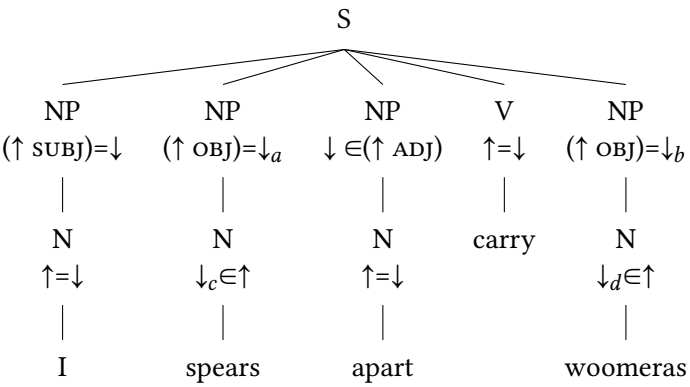
(27)
$$\left[\begin{array}{l} \text{INDEX} \left[\begin{array}{l}\text{PERS } 3\\ \text{NUM } \text{PL}\end{array}\right] \\ \left\{\left[\begin{array}{ll}\text{PRED} & \text{'GOANNA'} \\ \text{INDEX} & \left[\begin{array}{l}\text{PERS } 3\\ \text{NUM } \text{SG}\end{array}\right]\end{array}\right], \left[\begin{array}{ll}\text{PRED} & \text{'KANGAROO'} \\ \text{INDEX} & \left[\begin{array}{l}\text{PERS } 3\\ \text{NUM } \text{DU}\end{array}\right]\end{array}\right]\right\} \end{array}\right]$$

[8]This Yidiny example has been rewritten in a standard practical orthography which uses 'ny' for a palatal nasal, 'j' for a palatal stop and 'rr' for an alveolar trill.

Sadler & Nordlinger (2010) show how this approach to coordination also extends naturally to discontinuous examples such as (28) by combining with the standard LFG approach to discontinuity and non-configurationality discussed in Section 3.1 above. Since Economy of Expression allows all nodes to be optional unless independently required, each of the discontiguous coordinands can be represented at c-structure as a coordinate structure with just one daughter present (29), corresponding to the f-structure in (30).

(28) Kuuk Thaayorre
Ngul ngay kirk kempthe kal-m thul=yuk
then 1SG (ERG) spear(ACC) apart carry-PST.IPFV woomera(ACC)=STUFF
'I used to carry spears and woomeras separately' (Gaby 2006: 320)

(29)



(30)

$$\begin{bmatrix}
\text{PRED} & \text{'CARRY}\langle\text{SUBJ,OBJ}\rangle\text{'} \\
\text{ADJ} & \left\{\left[\text{PRED } \text{'APART'}\right]\right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 1 \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix}
\left\{\left(\begin{matrix} c \\ \end{matrix}\begin{bmatrix} \text{PRED} & \text{'SPEAR'} \\ \text{CASE} & \text{ACC} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{PL} \end{bmatrix} \end{bmatrix}\right), \left(\begin{matrix} d \\ \end{matrix}\begin{bmatrix} \text{PRED} & \text{'WOOMERA'} \\ \text{CASE} & \text{ACC} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{PL} \end{bmatrix} \end{bmatrix}\right)\right\} \\
\begin{matrix} a \\ b \end{matrix}\begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{PL} \end{bmatrix} \end{bmatrix}
\end{bmatrix}
\end{bmatrix}$$

All of the other instances of nominal juxtaposition exemplified above are also assumed to have the same syntactic structure with the differences between them arising from differences in the distribution of agreement features, and semantics. An appositional phrase such as (23), for example, is generated by the c-structure rule in (31), which is the same as the c-structure rule for coordination given in (25) except for the fact that each coordinand is associated with the appositional template @NP-APPOS instead of @NP-CNJT. The appositional template governs the distribution of agreement features as shown in (32). This ensures that the coordinated structure has the same INDEX features as each coordinand, as shown in the f-structure in (33).

(31) $\text{X} \longrightarrow \underset{\substack{\downarrow \in \uparrow \\ \text{@NP-APPOS}}}{\text{X}} , \underset{\substack{\downarrow \in \uparrow \\ \text{@NP-APPOS}}}{\text{X}}$

(32) NP-APPOS: $(\downarrow \text{IND}) \subseteq (\uparrow \text{IND})$

(33)
$$\begin{bmatrix} \text{INDEX} \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{SG} \\ \text{GEND} & \text{MASC} \end{bmatrix} \\ \left\{ \begin{bmatrix} \text{PRED} & \text{`HUSBAND'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{SG} \\ \text{GEND} & \text{MASC} \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{PRED} & \text{`OLD.MAN'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{SG} \\ \text{GEND} & \text{MASC} \end{bmatrix} \end{bmatrix} \right\} \end{bmatrix}$$

Sadler & Nordlinger (2010) show how the different juxtaposed structures can be captured in LFG by assuming that they all share the same syntactic structure (modulo differences in the distribution of agreement features, as illustrated above), while mapping onto different semantics. In this way the flexible architecture of LFG provides a unified account of a range of juxtaposed nominal constructions common to many Australian languages, while still accounting for their semantic differences, through the use of hybrid structures already motivated independently for analyses of coordination (Dalrymple & Kaplan 2000) (see also Patejuk forthcoming [this volume]).

3.3 Constructive case and case stacking

In Section 3.1 above we saw that case marking in non-configurational languages can encode grammatical relations, and saw that one way of capturing this in LFG is through the use of case conditionals. Nordlinger (1998b) provides an alternative approach, known as constructive case, which uses inside-out function application (see Belyaev forthcoming [this volume]) to capture the fact that the grammatical function information comes directly from the case morphology itself. Returning to the Warlpiri example discussed in (20a), on the constructive case approach the functional information associated with the ERG case would be that in (34):[9]

(34) (↑ CASE) = ERG
(SUBJ ↑)

The second line in this functional description specifies that the f-structure with which the case morphology is associated (i.e. ↑) is the value of a SUBJ function in a higher f-structure. Thus, the inflected nominal *kurdu-ngku* 'child-ERG' does not just encode the fact that the nominal is inflected with ergative case, but also that the nominal is functioning as a subject of the higher clause, corresponding to the f-structure given in (35):

(35) $\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'CHILD'} \\ \text{CASE} & \text{ERG} \end{bmatrix} \end{bmatrix}$

This approach has the benefit of capturing the essence of dependent-marking more accurately than the verb-mediated approaches described in Section 3.1 since the case-inflected nominal itself carries information about the grammatical function that it holds in the higher clausal f-structure. A further benefit, as discussed in detail by Nordlinger (1998b), is that it can straightforwardly capture other case behaviour found in dependent-marking Australian languages such as case stacking (Dench & Evans 1988; Andrews 1996), and the use of case morphology to mark clausal information such as tense/aspect/mood.

Case stacking arises through abundant case agreement, where a single nominal can carry multiple case markers, each one signalling a relationship to a higher level of structure. Consider the following examples:

[9]This is a slightly simplified representation for expository purposes. Nordlinger (1998b: 73) in fact suggests that the grammatical function information would be ((SUBJ ↑) OBJ) for ergative case, to capture the fact that it is only used with transitive subjects (i.e. subjects of f-structures that also contain an OBJ grammatical function).

(36) Warlpiri
Karnta-ngku ka=rla kurdu-ku miyi yi-nyi parraja-rla-ku.
woman-ERG PRS=3.DAT baby-DAT food.ABS give-NPST coolamon-LOC-DAT
'The woman is giving food to the baby (who is) in the coolamon.'
(Simpson 1991: 206)

(37) Martuthunira
Ngayu nhuwa-lhala tharnta-a kupuyu-marta-a thara-ngka-marta-a.
1.SG.NOM spear-PST euro-ACC little-PROP-ACC pouch-LOC-PROP-ACC
'I speared a euro with a little one in its pouch.' (Dench & Evans 1988: 7)

In (36) the locative-marked nominal *parraja-rla* 'coolamon' carries an additional case marker in agreement with the dative nominal *kurdu-ku* 'baby' which it modifies. Thus, the case marking on 'coolamon' specifies two different structural relationships: first, the locative case specifies that 'coolamon' functions as part of a locative adjunct, and then the dative case specifies that this locative adjunct is part of a higher dative-marked oblique argument. In (37), the most deeply embedded nominal *thara* 'pouch' is inflected with three case markers, each one specifying a successively higher structural relationship. Thus, the single inflected nominal *thara-ngka-marta-a* constructs the f-structure shown in (38).

(38)
$$\left[\text{OBJ}\left[\begin{array}{ll}\text{CASE} & \text{ACC}\\ \text{ADJ}_{\text{PROP}} & \left\{\left[\begin{array}{ll}\text{CASE} & \text{PROP}\\ \text{ADJ}_{\text{LOC}} & \left\{\left[\begin{array}{ll}\text{PRED} & \text{'POUCH'}\\ \text{CASE} & \text{LOC}\end{array}\right]\right\}\end{array}\right]\right\}\end{array}\right]\right]$$

Nordlinger (1998b) shows that this approach can account for a range of case stacking structures in Australian languages, as well as the interaction of case stacking with number marking and possession (see Chapters 4 and 5 therein). Sadler & Nordlinger (2004, 2006b) extend and improve Nordlinger's formal account to provide an analysis that integrates better with an LFG approach to the morphology-syntax interface (Sadler & Nordlinger 2004), and also show how Nordlinger's original morpheme-based account can be recast using a realizational approach to morphology (Sadler & Nordlinger 2006b). In some Australian languages, case morphology can also be used in complex clauses to encode cross-clausal reference and clause linkage relations. For discussion of how this use of case can be accounted for within the constructive case approach see Nordlinger (2000) and Austin (2016).

The fact that case morphology provides information to the clausal f-structure (by attributing a grammatical function to it) allows for case morphology to con-

tribute other types of clausal information as well, such as tense/aspect/mood.[10] Nordlinger (1998b: Chapter 4) shows that this is also found in some Australian languages, and can be accounted for straightforwardly with the constructive case approach. In Pitta Pitta (Blake 1987), for example, there are two ergative case morphs, one which is used in the future tense and the other in the non-future tense. The information associated with each of these can be represented as below, where the second f-description in each case specifies that the f-structure within which the case-marked nominal has a grammatical function (namely, the clausal f-structure) has a particular value for TENSE. The tense information associated with the case marker will be unified with the clausal f-structure and any tense information associated with the verb, thereby contributing to the overall tense value of the clause.

(39) *-lu* (↑ CASE) = ERG
((SUBJ ↑) TENSE) ≠ FUT

(40) *-ngu* (↑ CASE) = ERG
((SUBJ ↑) TENSE) = FUT

While it is typologically unusual for nominal morphology such as case to contribute clause-level information such as tense/aspect/mood, it is in fact found across languages of the world as shown by Nordlinger & Sadler (2004a,b). For a more detailed discussion of case in the LFG framework see Butt forthcoming [this volume].

3.4 Complex predicates

A number of Australian languages have complex predicates that take the form of light verb and coverb structures (see Andrews forthcoming(b) [this volume] for a more detailed discussion; the construction type focussed on here corresponds to type (1b) in this chapter). Detailed discussion of these constructions across Australian languages can be found in Schultze-Berndt (2000), McGregor (2000) and Bowern (2014). An example from Schultze-Berndt's discussion of Jaminjung is provided in (41).

(41) Jaminjung
walig gani-ma-m barrig
go.round 3SG:3SG-HIT-PRS paddock
'He walks around the fence (in a full circle).' (Schultze-Berndt 2000: 4)

[10] A different type of interaction between case morphology and the clause arises with semantic cases that can also function as clausal predicates – see Simpson (1991) for discussion.

In this construction the clausal predicate is formed through the combination of a finite inflected verb (e.g. *gani-ma-m*) with a coverb (e.g. *walig*). The two elements of the construction belong to distinct lexical classes, and thus are morphologically and syntactically different. Finite verbs are inflected for tense/aspect/mood and other verbal inflectional categories such as subject and object features. They form a closed class – in many languages restricted to between 10 and 30 members – and tend to have more general semantics (at least within the complex predicate). Coverbs, on the other hand, are usually uninflected, form a large open class and contribute more specific semantic content. The two elements together jointly determine the argument structure and event semantics, and therefore jointly construct the clausal predicate. In the languages of northern Australia where these constructions are found (see Bowern 2014), the majority of predicates are complex in this way.

Wilson (1999) provides a detailed LFG analysis of such complex predicates in Wagiman. Wilson shows that both the finite verb and the coverb in Wagiman are argument-taking predicates, and therefore each have their own PRED values, yet the complex predicate heads a single syntactic clause which in LFG requires a single clausal PRED at f-structure. To account for this, Wilson develops an account of complex predicate formation which uses a type of predicate fusion, modelled using lexical-conceptual structures (Jackendoff 1990), drawing on earlier work in LFG by Alsina (1993, 1996), Butt (1995, 1997), Mohanan (1994, 1997) and Andrews & Manning (1999).

Wilson's analysis follows that of Butt (1995, 1997) in using lexical conceptual structures (LCSs) to model complex predicate formation, but follows Andrews & Manning (1999) in locating these in f-structure (rather than a-structure as Butt does), replacing the PRED attribute with the more elaborated LCS attribute instead. Wilson proposes that the LCS of the coverb fuses into any position of the LCS of the finite verb where it is able to unify (Wilson 1999: 142). As an illustrative example, consider the complex predicate in (42):

(42) guk-ga nge-ge-na gahan warri-buga?
sleep-ASP 2SG-put-PST that child-PL
'Did you put the kids to sleep?' (Wilson 1999: 136)

According to Wilson's analysis, the finite verb *nge-ge-na* has the LCS in (43), and the coverb *guk-* has the LCS in (44).[11]

[11]The abbreviations used in the LCSs and associated attribute value matrices (AVMs) are as follows: the subscripted As denote positions which have to be linked to grammatical functions – in

(43) $[_{\text{Event}}\ \text{CAUSE}([_{\text{Thing}}\ ]_A, [_{\text{Event}}\ \text{BECOME}\ ([_{\text{State}}\ \text{BE}([_{\text{Thing}}\ ]_A, [_{\text{Place}}\ -])])])]$

(44) $[_{\text{State}}\ \text{BE}_{\text{Ident}}\ ([_{\text{Thing}}\ ]_A, [_{\text{Place}}\ \text{AT}_{\text{Ident}}\ ([_{\text{Property}}\ \textit{asleep}])])]$

These can be presented as attribute value matrices, as shown in (45) and (46) respectively.

(45) From Wilson (1999: 145: example (36)):

$$\begin{bmatrix} \text{TYPE} & \text{Event} \\ \text{FUNC} & \text{CAUSE} \\ \text{ARG1} & \begin{bmatrix} \text{TYPE} & \text{Thing} \\ \text{CONTENT} & \emptyset \\ \text{A-MARK} & \text{yes} \end{bmatrix} \\ \text{ARG2} & \begin{bmatrix} \text{TYPE} & \text{Event} \\ \text{FUNC} & \text{BECOME} \\ \text{ARG1} & \begin{bmatrix} \text{TYPE} & \text{State} \\ \text{FUNC} & \text{BE} \\ \text{ARG1} & \begin{bmatrix} \text{TYPE} & \text{Thing} \\ \text{CONTENT} & \emptyset \\ \text{A-MARK} & \text{yes} \end{bmatrix} \\ \text{ARG2} & [\text{TYPE}\ \ \text{Place}] \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

(46) From Wilson (1999: 147: example (39)):

$$\begin{bmatrix} \text{TYPE} & \text{State} \\ \text{FUNC} & \text{BE} \\ \text{FIELD} & \text{Ident} \\ \text{ARG1} & \begin{bmatrix} \text{TYPE} & \text{Thing} \\ \text{CONTENT} & \emptyset \\ \text{A-MARK} & \text{yes} \end{bmatrix} \\ \text{ARG2} & \begin{bmatrix} \text{TYPE} & \text{Place} \\ \text{FUNC} & \text{AT} \\ \text{FIELD} & \text{Ident} \\ \text{ARG1} & \begin{bmatrix} \text{TYPE} & \text{Property} \\ \text{CONTENT} & \text{asleep} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

the AVMs these correspond to the attribute A-MARK with the value 'yes'; 'Ident' stands for Identificational and is used to extend otherwise spatial functions (such as BE or AT) to the semantic field of ascription (thus, AT_{Ident} describes a property rather than a location); the value of the FUNC attribute in the AVMs is the function which expands the entity (e.g. GO, CAUSE, etc.). 'Thing' entities are not expanded by functions, but they can contain information about their referent, which is stored in the CONTENT attribute.

The c-structure rule which creates the complex predicate in (42) includes functional annotations that license and constrain predicate fusion through the unification of these LCSs. This is shown in (47), where C is the category 'coverb' (Wilson 1999: 144).

(47) $\overline{\text{V}} \longrightarrow \left(\begin{array}{c} \text{C} \\ \uparrow_{\text{LCS}}=\downarrow_{\text{LCS}} \\ (\uparrow \text{ LCS SF}^*)=(\downarrow \text{ LCS}) \end{array} \right), \begin{array}{c} \text{V} \\ \uparrow=\downarrow \end{array}$

The finite verb is annotated with ↑=↓ so that its inflectional features such as tense, aspect, and the information about the subject and object contribute to the f-structure of the complex predicate, and ultimately that of the clause. The annotations associated with the coverb ensure that (i) all information associated with the coverb apart from the LCS (e.g. any aspectual information) is contributed to the f-structure of the complex predicate, and (ii) the LCS of the coverb is fused into the LCS of the finite verb: (↑ LCS SF*) = (↓ LCS). Here SF stands for 'semantic function' and is defined as the set of attributes which can be contained in LCSs such as (45) and (46) (e.g. TYPE, FUNC, ARG1, ARG2). The use of functional uncertainty allows the LCS of the coverb – (↓ LCS) – to unify with any part of the LCS of the finite verb (the path consisting of any sequence of SFs, including none). So the f-structure will only be licit if the expansion of SF* picks out a place in the LCS of the finite verb where unification with the LCS of the coverb is possible. In the case of the complex predicate given in (42), based on the LCSs in (45) and (46), this path must be (↑ LCS ARG2 ARG1), since the coverb *guk* 'sleep' is of TYPE State, and there is only one place in the LCS of the finite verb where this can unify. Thus, the fused LCS for the complex predicate *guk -ge-* 'put to sleep' is that given in (48):

(48) From Wilson (1999: 147: example (39)):

$$\begin{bmatrix}
\text{TYPE} & \text{Event} & & & & & & \\
\text{FUNC} & \text{CAUSE} & & & & & & \\
\text{ARG1} & \begin{bmatrix} \text{TYPE} & \text{Thing} \\ \text{CONTENT} & \emptyset \\ \text{A-MARK} & \text{yes} \end{bmatrix} & & & & & & \\
\text{ARG2} & \begin{bmatrix} \text{TYPE} & \text{Event} & & & & & \\ \text{FUNC} & \text{BECOME} & & & & & \\ \text{ARG1} & \begin{bmatrix} \text{TYPE} & \text{State} & & & \\ \text{FUNC} & \text{BE} & & & \\ \text{FIELD} & \text{Ident} & & & \\ \text{ARG1} & \begin{bmatrix} \text{TYPE} & \text{Thing} \\ \text{CONTENT} & \emptyset \\ \text{A-MARK} & \text{yes} \end{bmatrix} & & \\ \text{ARG2} & \begin{bmatrix} \text{TYPE} & \text{Place} & & \\ \text{FUNC} & \text{AT} & & \\ \text{FIELD} & \text{Ident} & & \\ \text{ARG1} & \begin{bmatrix} \text{TYPE} & \text{Property} \\ \text{CONTENT} & \textit{asleep} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix} & & & & & & \\
\end{bmatrix}$$

Wilson shows that this approach to complex predicates in Wagiman can account for the range of different complex predicates found in the language, without requiring a radical extension of the LFG formalism beyond that already proposed by other complex predicate analyses (e.g. Alsina 1993, 1996, Butt 1995, 1997, Mohanan 1994, 1997 and Andrews & Manning 1999). This general approach to the formal analysis of complex predicate formation in Australian languages has also been adopted by Bowern (2004) for Bardi and Nordlinger (2010) for associated motion and motion serial verb constructions in Wambaya. An alternative approach to complex predicate formation using glue semantics as suggested in Andrews & Manning (1999) is proposed for the analysis of similar complex predicates in the central Australian language Arrernte by Dras et al. (2012).

Seiss (2013) provides a comprehensive analysis of the complex predicate system in Murrinhpatha which builds on the LCS-based approaches discussed above, but combines LCSs with a relational approach to lexical semantics, modeled with hierarchies of selectional restrictions. These hierarchies are then used to derive the argument structure of the complex predicates in the form of what Seiss calls LCS blueprints (based on the idea of templates, e.g. Dalrymple et al. (2004)). The blueprint LCS for causative complex predicates such as those in (49) and (50) is defined as in (51). The LCS blueprint states that the complex predicate expresses the meaning that something or someone (α) causes something (β) to become a certain result state with the help of some specific instrument. In Murrinhpatha

the complex predicate forms a single morphological word, and combines a classifier stem in first position in the verb, with a lexical stem (here *lerrkperrk*) in a subsequent position in the template. In a causative complex predicate, the result state is provided by the lexical stem while the instrument is provided by the classifier stem. For example, the lexical stem *lerrkperrk* 'crush' contributes the result state 'crushed', while the classifier stems 'do with HANDS' and 'do with FEET' contribute the instruments 'hand' and 'foot' respectively.

(49) Murrinhpatha
ku tumtum mam-lerrkperrk
CLF:ANIM egg 1SG.SBJ.HANDS.NFUT-crush
'I crushed the egg in my hand.' (Seiss 2013: 127)

(50) Murrinhpatha
ngunungam-lerrkperrk
1SG.SBJ.FEET.NFUT-crush
'I crushed the egg with my foot.' (Seiss 2013: 127)

(51) $$\begin{bmatrix} \text{CAUSE}\Big(\big[\text{Thing}\big]^{\alpha}_{A}, \big[\text{BECOME}\ \big(\big[\text{BE}\big(\big[\text{Thing}\big]^{\beta}_{A}, [\text{RESULT}]\big)\big]\big)\big]\Big) \\ \Big[\big[\text{BY}\ \big[\text{CAUSE}\Big(\big[\text{Thing}\big]^{\alpha}_{A}, \big[\text{AFF}^{-}\big([\text{INSTRUMENT}], \big[\text{Thing}\big]^{\alpha}_{A}\big)\big]\Big)\big]\big]\big] \end{bmatrix}$$

On this view, the classifier stem and the lexical stem do not each bring a complete LCS, but instead just a specific instrument (the classifier stem) or a specific result state (the lexical stem). The rest of the LCS is provided by the LCS blueprint. The lexical entries of the classifier stem and the lexical stem thus only consist of this information, as is illustrated in (52).

(52) do with HANDS: instrument = hand
do with FEET: instrument = foot
lerrkperrk: result = crushed

The LCS blueprint used by a particular combination is determined by the classifier and lexical stem together, whose compatibility is modelled by the hierarchies of selectional restrictions; the reader is referred to the comprehensive discussion in Seiss (2013) for further details. A notable aspect of Seiss's work on this topic is that, in addition to providing a comprehensive analysis of complex predicate combinations in Murrinhpatha, Seiss presents an implementation of Murrinhpatha's morphology using the Xerox finite-state technology tools XFST and LEXC

(Beesley & Karttunen 2003), and an implementation of some parts of Murrinhpatha's syntax using the XLE grammar development platform (Crouch et al. 2008).

Valence-changing constructions such as applicatives and causatives have also been analysed as complex predicates in many languages, including by Austin (1997), who draws on Alsina's (e.g. 1997) approach to complex predicates in analysing causatives and applicatives across a number of Australian languages (see Andrews forthcoming(b) [this volume] for further discussion of Austin's analysis in the context of LFG approaches to complex predicates).

4 Conclusion

In this chapter, I have covered the primary linguistic phenomena in Australian languages that have been given detailed analysis in LFG research, focussing particularly on the morphology-syntax interface, where the morphological complexity of Australian languages has made the most significant contributions to theoretical debate and development. Other areas where there has been some work on Australian languages, but for which space was not available for discussion here, include control and obviation constructions in Warlpiri (Simpson & Bresnan 1983), zero anaphora (**Austin2001**) and noun incorporation (Nordlinger & Sadler 2008; Baker & Nordlinger 2008; van Egmond 2008; Baker et al. 2010). Work on Australian languages within the LFG framework has also contributed to the discussion and analysis of grammatical relations cross-linguistically, in such areas as syntactic and morphological ergativity (Manning 1996), information structure and its role in case marking patterns (Simpson 2012), distinctions between syntactic and semantic cases (Andrews 2017) and the role of dative-marked NPs as core arguments or adjuncts (Simpson 1991). The majority of LFG researchers working on Australian languages are also descriptive linguists engaged in fieldwork and language documentation. This crossover has ensured that theoretical questions and implications arising from LFG analyses are fed back into language description work unearthing new findings about the languages and how they are structured, and ensuring that this research both contributes to the development of the LFG framework and to our understanding and description of these fascinating languages.

Acknowledgements

First and foremost, I would like to acknowledge all of the Indigenous language knowledge holders who have graciously worked with me and many other lin-

guists over the years to share their knowledge and teach us about their languages. Nothing discussed in this chapter would have been possible without their generosity and insight. For input and assistance with various aspects of this chapter, I would like to thank Peter Austin, Mary Laughren, Stephen Wilson, two anonymous referees, and especially Mary Dalrymple, editor extraordinaire, who is nothing short of a latex wizard and without whom this chapter would have been nothing but text! My research has been funded by the Australian Research Council over many years, most recently through the ARC Centre of Excellence for the Dynamics of Language (CE140100041), and I gratefully acknowledge their support.

References

Alsina, Alex. 1993. *Predicate composition: A theory of syntactic function alternations*. Stanford, CA: Stanford University. (Doctoral dissertation).

Alsina, Alex. 1996. *The role of argument structure in grammar: Evidence from Romance* (CSLI Lecture Notes). Stanford, CA: CSLI Publications.

Andrews, Avery D. 1996. Semantic case-stacking and inside-out unification. *Australian Journal of Linguistics* 16(1). 1–55. DOI: 10.1080/07268609608599530.

Andrews, Avery D. 2017. Case: Oblique, inherent, semantic, quirky. In Martin Everaert & Henk C. Riemsdijk (eds.), *The Wiley Blackwell companion to syntax*, 2nd edn., 1–36. Hoboken: John Wiley & Sons. DOI: 10.1002/9781118358733.wbsyncom086.

Andrews, Avery D. Forthcoming(a). Clause structure and configurationality. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

Andrews, Avery D. Forthcoming(b). Complex predicates. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

Andrews, Avery D. & Christopher D. Manning. 1999. *Complex predicates and information spreading in LFG*. Stanford, CA: CSLI Publications.

Austin, Peter K. 1997. Causatives and applicatives in Australian Aboriginal languages. In Kazuto Matsumura & Tooru Hayasi (eds.), *The dative and related phenomena*, 165–225. Tokyo: Hituzi Syobo.

Austin, Peter K. 2001. Word order in a free word order language: The case of Jiwarli. In Jane Simpson, David Nash, Mary Laughren, Peter K. Austin & Barry Alpher (eds.), *Forty years on: Ken Hale and Australian languages*, 305–323. Canberra: Pacific Linguistics. DOI: 10.15144/PL-512.

Austin, Peter K. 2016. Constructive morphology beyond case marking. Presentation at SE-LFG19 Workshop, SOAS.

Austin, Peter K. & Joan Bresnan. 1996. Non-configurationality in Australian aboriginal languages. *Natural Language & Linguistic Theory* 14. 215–268. DOI: 10.1007/bf00133684.

Baker, Brett, Kate Horrack, Rachel Nordlinger & Louisa Sadler. 2010. Putting it all together: Agreement, incorporation, coordination and external possession in Wubuy (Australia). In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 65–84. Stanford, CA: CSLI Publications.

Baker, Brett & Rachel Nordlinger. 2008. Noun-Adjective compounds in Gunwinyguan languages. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 109–128. Stanford, CA: CSLI Publications.

Beesley, Kenneth R. & Lauri Karttunen. 2003. *Finite state morphology*. Stanford, CA: CSLI Publications.

Belyaev, Oleg. Forthcoming. Core concepts of LFG. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

Blake, Barry J. 1987. *Australian Aboriginal grammar*. London: Croom Helm.

Blythe, Joseph. 2009. *Doing referring in Murrinh-Patha conversation*. Sydney: University of Sydney. (Doctoral dissertation).

Börjars, Kersti & John J. Lowe. Forthcoming. Noun phrases in LFG. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

Börjars, Kersti, Rachel Nordlinger & Louisa Sadler. 2019. *Lexical-Functional Grammar: An introduction*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/9781316756584.

Bowern, Claire. 2004. *Bardi verb morphology in historical perspective*. Canberra: Australian National University. (Doctoral dissertation). https://scholarship.rice.edu/bitstream/handle/1911/8322/Bowern_BVM.pdf.

Bowern, Claire. 2014. Complex predicates in Australian languages. In Harold Koch & Rachel Nordlinger (eds.), *The languages and linguistics of Australia: A comprehensive guide*, 263–294. Berlin: De Gruyter.

Bowern, Claire (ed.). 2021. *The Oxford guide to Australian languages*. Oxford: Oxford University Press. Forthcoming.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

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. 1987. Topic, pronoun, and agreement in Chicheŵa. *Language* 63. 741–782. DOI: 10.2307/415717.

Butt, Miriam. 1995. *The structure of complex predicates in Urdu* (Dissertations in Linguistics). Stanford, CA: CSLI Publications.

Butt, Miriam. 1997. Complex predicates in Urdu. In Alex Alsina, Joan Bresnan & Peter Sells (eds.), *Complex predicates*, 107–150. Stanford, CA: CSLI Publications.

Butt, Miriam. Forthcoming. Case. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

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.

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.

Dench, Alan & Nicholas Evans. 1988. Multiple case-marking in Australian languages. *Australian Journal of Linguistics* 8(1). 1–47. DOI: 10.1080/07268608808599390.

DITRC, Australian Institute of Aboriginal and Torres Strait Islander Studies, Jacqueline Battin, Jason Lee, Douglas Marmion, Rhonda Smith, Tandee Wang, Australian National University, Yonatan Dinku, Janet Hunt, Francis Markham, Denise Angelo, Emma Browne, Inge Kral, Carmel O'Shannessy, Jane Simpson & Hilary Smith. 2020. *National indigenous languages report*. Tech. rep. Canberra: Department of Infrastructure, Transport, Regional Development, Communications, formerly the Department of Communications & the Arts. http://www.arts.gov.au/what-we-do/indigenous-arts-and-languages/national-indigenous-languages-report.

Dixon, R. M. W. 1977. *A grammar of Yidiny*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/cbo9781139085045.

Dixon, R. M. W. 2002. *Australian languages: Their nature and development*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/cbo9780511486869.

Dras, Mark, Benjamin Borschinger, Yasaman Motazedi, Myfany Turpin, Francois Lareau, Robert Dale, Owen Rambow & Morgan Ulinski. 2012. Complex predicates in Arrernte. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '12 conference*, 177–197. Stanford, CA: CSLI Publications.

Evans, Nicholas. 1995. *A grammar of Kayardild: with historical-comparative notes on Tangkic*. Berlin: Walter de Gruyter. DOI: 10.1515/9783110873733.

Evans, Nicholas. 2003. *Bininj Gun-wok: A pan-dialectal grammar of Mayali, Kunwinjku and Kune*. Canberra: Pacific Linguistics. http://hdl.handle.net/1885/53188.

Evans, Nicholas. 2007. Warramurrungunji undone: Australian languages in the 51st millennium. In Matthias Brenzinger (ed.), *Language diversity endangered*, 343–373. Berlin: Mouton de Gruyter.

Evans, Nicholas & Rachel Nordlinger. 2004. Extreme morphological shift: Verbal case in Kayardild. Presented at the LFG '04 Conference.

Evans, Nicholas & Hans-Jürgen Sasse. 2002. Introduction: Problems of polysynthesis. In Nicholas Evans & Hans-Jürgen Sasse (eds.), *Problems of polysynthesis*, 1–13. Berlin: Akademie Verlag.

Fletcher, Janet & Andrew Butcher. 2014. Sound patterns of Australian languages. In Harold Koch & Rachel Nordlinger (eds.), *The languages and linguistics of Australia: A comprehensive guide*, 91–138. Berlin: De Gruyter.

Gaby, Alice. 2006. *A grammar of Kuuk Thaayorre*. Melbourne: University of Melbourne. (Doctoral dissertation). http://repository.unimelb.edu.au/10187/392.

Hale, Ken. 1981. *On the position of Warlpiri in a typology of the base*. Bloomington, IN: Indiana University Linguistics Club. Distributed by Indiana University Linguistics Club, Bloomington.

Hale, Ken. 1982. Preliminary remarks on configurationality. In James Pustejovsky & Peter Sells (eds.), *Proceedings of the 12th annual meeting of the North Eastern Linguistic Society*, 86–96. Amherst, MA: GLSA.

Hale, Ken. 1983. Warlpiri and the grammar of non-configurational languages. *Natural Language & Linguistic Theory* 1. 5–47. DOI: 10.1007/bf00210374.

Haug, Dag. Forthcoming. Agreement. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

Jackendoff, Ray. 1990. *Semantic structures*. Cambridge, MA: The MIT Press.

Koch, Harold & Rachel Nordlinger (eds.). 2014a. *The languages and linguistics of Australia: A comprehensive guide*. Berlin: De Gruyter. DOI: 10.1515/9783110279771.

Koch, Harold & Rachel Nordlinger. 2014b. The languages of Australia in linguistic research: context and issues. In Harold Koch & Rachel Nordlinger (eds.), *The languages and linguistics of Australia: A comprehensive guide*, 3–21. Berlin: De Gruyter.

Louagie, Dana. 2020. *Noun Phrases in Australian languages*. Berlin: De Gruyter.

Louagie, Dana & Jean-Christophe Verstraete. 2016. Noun phrase constituency in Australian languages: a typological study. *Linguistic Typology* 20(1). 25–80. DOI: 10.1515/lingty-2016-0002.

Manning, Christopher D. 1996. *Ergativity: Argument structure and grammatical relations* (Dissertations in Linguistics). Stanford, CA: CSLI Publications.

Mansfield, John. 2019. *Murrinhpatha morphology and phonology*. Berlin: Mouton de Gruyter.

McGregor, William B. 2000. *Verb classification in Australian languages*. Berlin: Mouton de Gruyter. DOI: 10.1515/9783110870879.

Meakins, Felicity & Rachel Nordlinger. 2014. *A grammar of Bilinarra: An Australian aboriginal language of the Northern Territory*. Berlin: De Gruyter Mouton. DOI: 10.1515/9781614512745.

Mohanan, Tara. 1994. *Argument structure in Hindi* (Dissertations in Linguistics). Stanford, CA: CSLI Publications.

Mohanan, Tara. 1997. Multidimensionality of representations: NV complex predicates in Hindi. In Alex Alsina, Joan Bresnan & Peter Sells (eds.), *Complex predicates*, 431–472. Stanford, CA: CSLI Publications.

Mujkic, Elena. 2013. *Murrinh-Patha syntax: Clausal structure and the noun phrase*. Melbourne: University of Melbourne. (Honours thesis).

Nichols, Johanna. 1986. Head-marking and dependent-marking grammar. *Language* 62(1). 56–119. DOI: 10.1353/lan.1986.0014.

Nordlinger, Rachel. 1998a. *A grammar of Wambaya, Northern Australia*. Canberra: Pacific Linguistics.

Nordlinger, Rachel. 1998b. *Constructive case: Evidence from Australian languages*. Stanford, CA: CSLI Publications.

Nordlinger, Rachel. 2000. Australian case systems: Towards a constructive solution. In Miriam Butt & Tracy Holloway King (eds.), *Argument realization*, 41–72. Stanford, CA: CSLI Publications.

Nordlinger, Rachel. 2010. Complex predicates in Wambaya: Detaching predicate composition from syntactic structure. In Mengistu Amberber, Brett Baker & Mark Harvey (eds.), *Complex predicates: Cross-linguistic perspectives on event Structure*, 237–258. Cambridge, UK: Cambridge University Press.

Nordlinger, Rachel. 2017. The languages of the Daly River region (Northern Australia). In Michael Fortescue, Marianne Mithun & Nicholas Evans (eds.), *The Oxford handbook of polysynthesis*, 782–807. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199683208.013.44.

Nordlinger, Rachel. Forthcoming. LFG and Australian languages. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

Nordlinger, Rachel & Joan Bresnan. 1996. Nonconfigurational tense in Wambaya. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '96 conference*, 338–352. Stanford, CA: CSLI Publications.

Nordlinger, Rachel & Joan Bresnan. 2011. Lexical-functional grammar: Interactions between morphology and syntax. In Robert D. Borsley & Kersti Börjars (eds.), *Non-transformational syntax: Formal and explicit models of grammar*, 112–140. Oxford: Wiley-Blackwell. DOI: 10.1002/9781444395037.ch3.

Nordlinger, Rachel & Louisa Sadler. 2004a. Nominal tense in cross-linguistic perspective. *Language* 80(4). 776–806. DOI: 10.1353/lan.2004.0219.

Nordlinger, Rachel & Louisa Sadler. 2004b. 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. 2008. From juxtaposition to incorporation: An approach to generic-specific constructions. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 394–412. Stanford, CA: CSLI Publications.

Patejuk, Agnieszka. Forthcoming. Coordination. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

Reinöhl, Uta. 2020. Continuous and discontinuous nominal expressions in flexible (or "free") word order languages. *Linguistic Typology* 24. 71–111.

Sadler, Louisa & Rachel Nordlinger. 2004. Relating morphology to syntax. In Louisa Sadler & Andrew Spencer (eds.), *Projecting morphology*, 159–186. Stanford, CA: CSLI Publications.

Sadler, Louisa & Rachel Nordlinger. 2006a. Apposition as coordination: Evidence from Australian languages. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '06 conference*, 437–454. Stanford, CA: CSLI Publications.

Sadler, Louisa & Rachel Nordlinger. 2006b. Case stacking in realizational morphology. *Linguistics* 44(3). 459–487. DOI: 10.1515/ling.2006.016.

Sadler, Louisa & Rachel Nordlinger. 2010. Nominal juxtaposition in Australian languages: An LFG analysis. *Journal of Linguistics* 46(2). 415–452.

Schultze-Berndt, Eva. 2000. *Simple and complex verbs in Jaminjung: A study of event categorization in an Australian language*. Nijmegen: Katholieke Universiteit Nijmegen & Max Planck Institute for Psycholinguistics. (Doctoral dissertation).

Seiss, Melanie. 2013. *Murrinh-Patha complex verbs: Syntactic theory and computational implementation*. Konstanz: University of Konstanz. (Doctoral dissertation). http://nbn-resolving.de/urn:nbn:de:bsz:352-0-308493.

Seiss, Melanie & Rachel Nordlinger. 2010. Applicativizing complex predicates: A case study from Murrinh-Patha. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '10 conference*, 416–436. Stanford, CA: CSLI Publications.

Sharp, Janet. 2004. *Nyangumarta: A language of the Pilbara region of Western Australia*. Canberra: Pacific Linguistics. http://hdl.handle.net/1885/146175.

Simpson, Jane. 1983. *Aspects of Warlpiri morphology and syntax*. Cambridge, MA: Massachusetts Institute of Technology. (Doctoral dissertation). https://dspace.mit.edu/handle/1721.1/15468. Published in revised form as Simpson (1991).

Simpson, Jane. 1991. *Warlpiri morpho-syntax: A lexicalist approach*. Dordrecht: Kluwer Academic Publishers.

Simpson, Jane. 2007. Expressing pragmatic constraints on word order in Warlpiri. 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*. Stanford, CA: CSLI Publications.

Simpson, Jane. 2012. Information structure, variation and the referential hierarchy. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts & Paul Trilsbeek (eds.), *Potentials of language documentation: Methods, analyses, and utilization*, 83–89. Honolulu: University of Hawai'i Press.

Simpson, Jane & Joan Bresnan. 1983. Control and obviation in Warlpiri. *Natural Language & Linguistic Theory* 1(1). 49–64. DOI: 10.1007/bf00210375.

Singer, Ruth. 2018. A small speech community with many small languages: the role of receptive multilingualism in supporting linguistic diversity at Warruwi Community (Australia). *Language & Communication* 62. 102–118. DOI: 10.1016/j.langcom.2018.05.002.

Snijders, Liselotte. 2015. *The nature of configurationality in LFG*. Oxford: University of Oxford. (D.Phil. Thesis).

Snijders, Liselotte. 2016. An LFG account of discontinuous nominal expressions. In *Proceedings of DiscoNLP*, 1–11. DOI: 10.18653/v1/w16-0901.

Toivonen, Ida. Forthcoming. Pronoun incorporation. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.

van Egmond, Marie-Elaine. 2008. Incorporated adjunct classifiers in Anindilyakwa: An empirical challenge to LFG. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '08 conference*, 235–251. Stanford, CA: CSLI Publications.

Wilson, Stephen. 1999. *Coverbs and complex predicates in Wagiman*. Stanford, CA: CSLI Publications.

Handbook of Lexical Functional Grammar

Set blurb on back with \BackBody{my blurb}