Information Structure in Topological Dependency Grammar

Geert-Jan Kruijff
Computational Linguistics
Saarland University
Saarbrücken, Germany
`gj@coli.uni-sb.de`

Denys Duchier
INRIA Lorraine
Nancy, France
`denys.duchier@loria.fr`

Abstract

Topological Dependency Grammar (TDG) is a lexicalized dependency grammar formalism, able to model languages with a relatively free word order. In such languages, word order variation often has an important function: the realization of information structure. The paper discusses how to integrate information structure into TDG, and presents a constraint-based approach to modelling information structure and the various means to realize it, focusing on (possibly simultaneous use of) word order and tune.

1 Introduction

In this paper, we present an extension to Topological Dependency Grammar (Duchier and Debusmann, 2001) enabling us to analyse e.g. word order variation and tune as means to indicate what is the *topic* and what is the *focus* of an expression – i.e. its information structure (cf. §2, §4). Using a constraint-based approach, we can analyse the surface form of an expression in terms of the information structure that it realizes.

The information structure of an expression is a core part of its meaning: it indicates how the expression relates to the discourse context. Information structure thus constitutes a crucial factor in determining an expression's contextual appropriateness or interpretability. Particularly in applications that involve human-computer interaction, information structure has thus been found to have a great impact on the understandability of computer-generated language, e.g. question/answering dia-

logues (Prevost and Steedman, 1994; Hoffman, 1995; Kruijff-Korbayová et al., 2003) or generation (Kruijff-Korbayová et al., 2002).

In this paper we concentrate on information structure and the syntax/semantics-interface: We want to be able to reconstruct an expression's information structure at the level of meaning, given the expression's surface form.

To realize information structure a language may employ a variety of means, not only word order or tune but also morphology or marked syntactic constructions. Collectively we call these means *structural indications of informativity*, after (Vallduví and Engdahl, 1996; Kruijff, 2001).

As §2 illustrates, languages are not restricted to using just a single means. Within a single expression several types of indications can normally be used simultaneously. The indications may constrain the expression's well-formedness, and it is through their interaction that the indications help realize information structure.

It is precisely this interaction that presents a problem for existing accounts of information structure and its realization. Although accounts normally acknowledge that there are various types of structural indications, most of them focus solely on modelling the use of a single type of structural indication. For example, (Steedman, 2000) focuses on tune, (Hoffman, 1995) or (Hajičová et al., 1995) focus on word order.

Such focus would be unproblematic if it were clear how these accounts could be extended to cover multiple, interacting types of structural indications. However, even for (Steedman, 2000; Hoffman, 1995), which are the formally most detailed, this is by no means obvious. CCG's underlying principles (notably, the *Principle of Ad-*

jacency) forces Hoffman to introduce separate derivations for establishing an expression's syntactic structure (incl. word order) and its information structure. This detaches information structure from word order as an indication of the former, a problem that arguably gets aggravated if one were to try to incorporate Steedman's model of tune.

The contribution we make here is the presentation of a framework that (i) can describe the use of any number of structural indications in realizing information structure in a perspicuous way, and that (ii) is amenable to a formalization in the style of TDG to extend the latter's efficient constraint-based parser. A proviso: Given the limited space, we do not deal with contrast in this paper.

Overview: §2 presents data motivating our point that languages can use several types of structural indications of informativity simultaneously, and the effect this may have on grammaticality. §3 introduces the necessary basic concepts of TDG. In §4 we discuss how to extend TDG to deal with information structure: We outline the underlying linguistic model, and specify the formal details of the extension. The resulting model we then apply to the data of §2. We close with conclusions.

2 Motivation

When a speaker wants to communicate some meaning to a hearer, she does that against a background of discourse referents that have already been activated in the context, and which are (presumably) shared between speaker and hearer. The meaning a speaker communicates relates to these already established referents, and presents more ("new") information *about* these referents. The former part of the meaning we call the *topic*, the latter the *focus*. An expression's information structure is the division of its meaning into a topic and a focus (Sgall et al., 1986; Vallduví, 1990).

Languages may realize information structure in various ways. For example, in a language with a relatively free word order, variations in linearization are prototypically used to indicate different information structure (Sgall et al., 1986; Hoffman, 1995; Kruijff, 2001). This explains why different variations, though equally grammatical, are usually not equally interchangeable in a given context.

To illustrate the idea of context-dependence, consider the Czech example in (1) and its grammatical variations in (2).[1]

(1) [Snědl]$_F$ [Honza]$_F$ [koblihu]$_F$.
eat-PAST John   donut
"John ate a donut."

(2)   a. [Honza]$_T$ snědl [koblihu]$_F$.
    b. [Koblihu]$_T$ snědl [Honza]$_F$.
    c. [Honza koblihu]$_T$ [snědl]$_F$.

(1) illustrates an "all-focus" sentence – the entire meaning is new. The examples in (2) presuppose different items to be present ("salient") in the already established dialogue. For example, if the speaker utters (2b) in a context where there is no donut, the hearer would most likely reply with "What donut?!", whereas (2a) assumes that "Honza" is a person the hearer can identify.

Not every language has a relatively free word order, though. English has a fixed word order where it concerns complements, and therefore usually resorts to using tune to realize information structure. The examples in (3) illustrate several possible information structures, given the placement of the pitch accent.[2]

(3)   a. [John]$_F$ gave [Mary]$_F$ ["Moby DICK"]$_F$.
    b. [JOHN]$_F$ [gave]$_T$ [Mary]$_T$ ["Moby Dick"]$_T$.
    c. [John]$_T$ [gave]$_T$ [MARY]$_F$ ["Moby Dick"]$_T$.

Particularly in languages that have a degree of word order freedom inbetween English and Slavic languages like Czech, we can find examples of a strong interaction between word order and tune. For example, consider the Dutch examples in (4) and (5). (4) illustrates the all-focus case. (5a–c) show well-formed variations interpretable on different contexts. (5d) however, is ill-formed. By placing "Moby Dick" sentence initial and putting a non-contrastive stress on it, it gets interpreted as the subject of the (active) verb "lezen".

(4) Jan las   "Moby   DICK"
John read-PAST "Moby Dick"
John read "Moby DICK"

(5)   a. (Who read "Moby Dick"?)
    [JAN]$_F$ [las]$_T$ ["Moby Dick"]$_T$.
    "JOHN read "Moby Dick"."
    b. (Who read "Moby Dick"?)
    ["Moby Dick"]$_T$ [las]$_T$ [JAN]$_F$.
    "JOHN read "Moby Dick"."

---

[1] Subscript T indicates that the item belongs to the topic, F that it belongs to the focus.
[2] SMALL CAPS indicate pitch accent.

c. (What did John read?)
[Jan]$_T$ [las]$_T$ ["MOBY DICK"]$_F$.
"John read "MOBY DICK"."

d. (What did John read?)
* ["MOBY DICK"]$_F$ [las]$_T$ [Jan]$_T$.
*"MOBY DICK" read John.

We would like to argue that similar interactions between word order and tune can also be observed in English. English has more freedom in ordering adjuncts, as (7) illustrates (Sgall et al., 1986). (6) presents the all-focus case.

(6) John flew from London to Paris on Tuesday.

(7) a. [On Tuesday]$_T$, [John]$_F$ [flew]$_F$ [from London]$_F$ [TO PARIS]$_F$.

b. [On Tuesday]$_T$, [John]$_T$ [flew]$_T$ [to Paris]$_T$ [FROM LONDON]$_F$.

c. [From London]$_T$, [John]$_T$ [flew]$_T$ [to Paris]$_T$ [ON TUESDAY]$_F$.

The boundaries between topic and focus in (7) arise from non-canonical ordering of adjuncts, and the tendency of SVO languages like English to place focus items towards the end of the sentence. For example, in (7b) the to-PP and from-PP are inverted – the from-PP is part of the focus, whereas the non-canonical ordering of the to-PP and the from-PP makes us place the topic/focus-boundary between these two PPs. The same idea applies to (7a): Only the on-PP is ordered non-canonically with respect to the rest of the complements and adjuncts, hence we put the topic/focus-boundary between the on-PP and the subject. We elaborate this in §4.

English is relatively free in placing pitch accent – given a canonical order, (3). When varying the word order as in (7), we find that the interaction between word order and tune leads to strong preferences in interpretation.[3] The examples in (8) illustrate this effect. We interpret elements from the question as topical (in the answer).

(8) On Tuesday, what flight did John take?

a. [On Tuesday]$_T$, [John]$_T$ [flew]$_T$ [from London]$_F$ [TO PARIS]$_F$.

b. ?#[On Tuesday]$_T$, [John]$_T$ [flew]$_T$ [to Paris]$_F$ [FROM LONDON]$_F$.

c. #[On Tuesday]$_T$, [John]$_T$ [flew]$_T$ [TO PARIS]$_F$ [from London]$_T$.

---

[3]We see these preferences as a weaker version of the effect such interaction has on well-formedness observed for Dutch.

The example in (8c) leads to a dispreferred (#) interpretation: In English, constituents coming after the pitch accent (here, TO PARIS) are interpreted by default as given (resulting in ["from London"]$_T$). Though the word order is well-formed, as is the placement of the pitch accent on the to-PP, the resulting surface form is not appropriate in the given context. (8b) is #'d because its non-canonical ordering of the PPs would suggest a topic/focus-boundary between the to-PP and the FROM-PP, suggesting the to-PP to be given.[4]

To recapitulate, variation in the placement of (non-contrastive) pitch accent or in word order helps indicate the boundary between topic and focus. Furthermore, when tune and word order are both used to realize information structure, they constrain one another. In §4 we present a formalization in TDG that captures these phenomena. Before that, we use the next section to present the necessary basics of TDG.

3 Topological Dependency Grammar

Duchier and Debusmann (2001) introduced TDG, a lexicalized formalism for dependency grammar, to tackle linearization phenomena in freer word-order languages. These are explained as emerging from the interaction of a non-ordered tree of syntactic dependencies, where edges are labeled by grammatical functions, with an ordered and a projective tree of topological dependencies, where edges are labeled by topological fields. Both trees are simultaneously constrained by a lexical assignment that e.g. restricts the licensed edges. Furthermore TDG stipulates that they must be related by an emancipation mechanism whereby a word is allowed to *climb up* and *land* in the topological domain of a syntactic ancestor.

For example, the German sentence

(9) Maria überredet ihn ein Buch zu lesen
Mary convinces him a book to read

receives the following analysis, where (10) is the syntax tree and (11) the topological tree:

---

[4]Native speakers prefer (8a) over (8b), yet do not rule out (8b) as strongly as (8c); hence the ? with (8b).

(10)

subj · obj · vinf · obj

Maria überredet ihn ein Buch zu lesen

(11)

vf · v12 · mf · mf · vcf

n · n · n · v

Maria überredet ihn ein Buch zu lesen
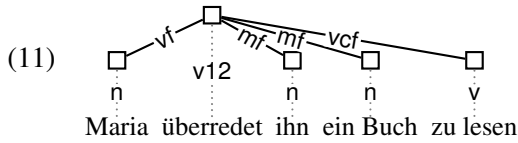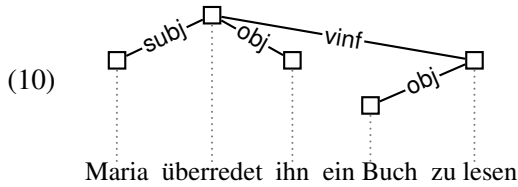
Notice that, while "Buch" is the syntactic object of "lesen" it lands in the Mittelfeld (mf) of the main verb "überredet".

On-going work on the development of a syntax/semantics interface for TDG extends the same methodology to the recovery of deep semantic dependencies. An additional structure is introduced: the semantic argument structure. This is a directed acyclic graph with edges labeled by semantic relations. For sentence (9) above, the corresponding argument structure is given in (12):

(12)

actor · patient · purpose · actor · patient

Maria überredet ihn ein Buch zu lesen
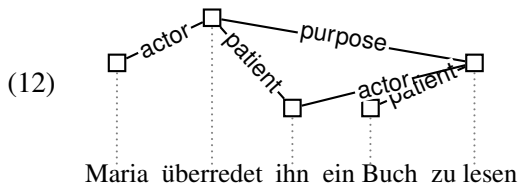
Notice that "ihn" is now both the patient of "überredet" and the actor of "lesen". Again, TDG postulates an emancipation mechanism relating the argument structure to the syntax tree, that e.g. allows a (subject) semantic dependent to *climb up* and be realized as a raised syntactic argument of a dominating control or raising verb.

In the present paper, we take advantage of this extension to TDG, and avail ourselves of the argument structure. For more details on how TDG can model word order, we refer to Duchier and Debusmann (2001).

4 Modelling information structure realization

The goal of the current section is to present a TDG-based model of how word order and intonation may together help realize information structure. In §4.1 we present the linguistic theory our model formalizes, after which we present the formalization itself in §4.2. We apply the model to various examples from §2 in §4.3.

4.1 Linguistic background

Some theories define topic and focus as *atomic* terms, often corresponding to a concrete division of an expression's surface form, e.g. (Vallduví, 1990). Here, we take a more recursive perspective, like (Sgall et al., 1986; Hajičová et al., 1998; Steedman, 2000): topic and focus are established (recursively) on the basis of the informativity of individual *(discourse) referents* that make up an expression's meaning. If the speaker presents a referent as activated in the preceding context (association/direct introduction), then we call that referent *contextually bound* (CB). If a referent has not been activated yet, we call it *contextually nonbound* (NB).

Decoupling the definition of topic and focus from surface realization and defining them recursively enables us to deal in a perspicuous way with discontinuous topics/foci and embedding.

There are numerous sources providing indications of whether a referent is CB or NB: contextual activation, lexical semantics, variations in word order, tune, morphology, etc. The challenge is to meaningfully combine them. In this paper, we consider a simple approach based on the classical 4-valued Boolean lattice: ⊤ is the top of the lattice and indicates the absence of information, CB and NB are the two boolean options, and ⊥ represents a contradiction. Such an approach is well-suited for integration into TDG's constraint-based framework. Below we describe several principles that derive indications of CB/NB-ness in the form of values in the Boolean lattice. Their conclusions are then combined by ⊓ (lattice meet) for contribution to the expression's information structure.

Contextual activation. If a discourse referent is activated in the preceding context either through association or direct introduction, then it is assigned CB, else ⊤.

Tune. Tune is another source of partial information about CB/NB-ness. We assume that a pitch accent indicates NB. Following (Steedman, 2000), we assume that CB is assigned to the siblings (or

dependents, if the verb has a pitch accent) rightward of the pitch accent. Otherwise, we assign $\top$.

Lexical semantics. Lexical semantics may also provide indications about CB/NB-ness. For example, in the simplified setting of this paper, we assume that the English indefinite article "a" prototypically indicates NB, while the definite article "the" indicates CB. In other cases, lexical semantics simply assigns $\top$.

Systemic ordering. Like Sgall et al. (1986), we assume that there is a canonical ordering over dependents such as ACTOR, PATIENT, LOCATION etc, and that variation on this order indicates differences in informativity (cf. the examples in (7)). We call this order the *systemic ordering* (SO), and allow each verb to have its own lexicalized SO.[5] The SO for many English verbal dependents is:

(13) ACTOR < ADDRESSEE < PATIENT < FROMWHERE < WHERETO < TIMEWHEN

SO relates to CB/NB-ness as follows. For SVO and OV languages, we assume that the trailing sequence of verbal dependents that are realized in canonical order at the clause level are assigned $\top$, while all preceding ones are considered to be CB. Thus we are mostly interested in the rightmost violation of SO among the dependents of a given verbal head. For example, given the SO of (13), we can explain why "Tuesday" is CB in (14): Its actual linearization is non-canonical wrt. the SO, while all following dependents of the clause *are* linearized in canonical order.

(14) On Tuesday$_{\text{CB}}$, John$_\top$ flew$_\top$ from London$_\top$ TO PARIS$_\top$.

Projection. It is possible that no source of information determines the NB/CB-ness of a particular word. In this case, the principle of *projection* enables us to extend an assignment starting from a referent whose NB/CB-ness is known:

For SVO and OV languages, if a referent δ is NB, then referents left of δ can also be considered NB (*projection*) if they are (incl. δ) ordered canonically wrt. SO and are not already determined to be CB. CB-ness can project leftwards

---

[5](Sgall et al., 1986) posit SO as a universal order, holding equally across all verbs. However, that seems to contradict the results in (Kurz et al., 2000).

over referents ordered either canonically or non-canonically wrt. SO.

For example, consider (15).

(15)   a. John gave Mary a book TODAY$_{\text{NB}}$.
       b. John gave Mary a book$_{\text{NB}}$ TODAY$_{\text{NB}}$.
       c. John gave Mary$_{\text{NB}}$ a book$_{\text{NB}}$ TODAY$_{\text{NB}}$.
       $\vdots$

All dependents in (15) are ordered canonically wrt. SO. Hence, when the pitch accent on "today" specifies it as NB, we can project NB-ness leftwards over all the preceding referents (resulting in an all-focus sentence). If we would have "the book" instead, we could not project NB-ness. Instead, we could project CB leftwards from "the book".

In the next sections we formalize and illustrate the principles on examples involving indications following from all of the factors mentioned above: Word order, tune, lexical semantics, projection, and contextual activation.

4.2 Formalization in TDG

In this section, we outline how the model theoretic approach of TDG (Duchier, 2001) can be extended in the same spirit with a formalization of systemic order violations, thus setting the stage for a contraint-based account of information structure.

We write $\mathcal{E}$ for the set of lexical entries, i.e. the lexicon, and $\mathcal{L}_{\text{TH}}$ for the set of semantic dependency relations. Each lexical entry stipulates a systemic ordering on $\mathcal{L}_{\text{TH}}$, which we model using the function:

$$\text{so} : \mathcal{E} \to \mathcal{L}_{\text{TH}} \times \mathcal{L}_{\text{TH}}$$

Given a lexical assignment $\alpha : V \to \mathcal{E}$ of lexical entries to the words V of a sentence, we overload the function as follows to obtain the systemic order lexically assigned to each word $w \in V$:

$$\text{so}(w) = \text{so}(\alpha(w))$$

The semantic argument structure (V, E_{TH}) is a DAG with edges $E_{\text{TH}} \subseteq V \times V \times \mathcal{L}_{\text{TH}}$. Each semantic role θ can also be interpreted as a function from words to sets of words:

$$\theta(w) = \{w' \in V \mid (w, w', \theta) \in E_{\text{TH}}\}$$

In this paper, we assume that each $\theta(w)$ contains at most one element and that for any $\theta_1 \neq \theta_2 \in \mathcal{L}_{\mathrm{TH}}$, $\theta_1(w) \cap \theta_2(w) = \emptyset$, i.e. that the semantic arguments of one head are all distinct.

Given $\mathsf{so}(w)$ we can define the systemic order $\mathsf{so{:}args}(w) \subseteq V \times V$ induced on w's actual semantic dependents:

$$\mathsf{so{:}args}(w) = \cup\{\theta_1(w) \times \theta_2(w) \mid (\theta_1, \theta_2) \in \mathsf{so}(w)\}$$

The topological structure, which is part of a TDG analysis, provides us with a total order $\prec$ on V. We write $\mathcal{L}_{\mathrm{TH}}(w) = \cup\{\theta(w) \mid \theta \in \mathcal{L}_{\mathrm{TH}}\}$ for the set of w's semantic dependents and $\prec|_{\mathcal{L}_{\mathrm{TH}}(w)}$ for the restriction of $\prec$ to $\mathcal{L}_{\mathrm{TH}}(w)$.

The set $\mathsf{nso{:}args}(w)$ of non-systematically ordered pairs of w's semantic dependents can be obtained by the following set difference:

$$\mathsf{nso{:}args}(w) = \prec|_{\mathcal{L}_{\mathrm{TH}}(w)} \setminus \mathsf{so{:}args}(w)$$

we wish to identify the set of all semantic dependents of w that either violate systemic order or are left of one that does. Given an ordering R, we write $\mathsf{dom}(R)$ for its underlying domain, $\pi_1(R)$ resp. $\pi_2(R)$ for its 1st resp. 2nd projections, and $\mathsf{eqleft}(w)_R$ for the set of elements left of or equal to w in R:

$$\pi_1(R) = \{x \mid (x, y) \in R\}$$
$$\pi_2(R) = \{y \mid (x, y) \in R\}$$
$$\mathsf{dom}(R) = \pi_1(R) \cup \pi_2(R)$$
$$\mathsf{eqleft}(w)_R = \{w\} \cup \{w' \mid (w', w) \in R\}$$

Thus the set of dependents to be assigned CB according to the systemic ordering principle is:

$$\mathcal{L}_{\mathrm{TH}}(w) \cap$$
$$\cup\{\mathsf{eqleft}(w')_\prec \mid w' \in \pi_1(\mathsf{nso{:}args}(w))\}$$

Other principles, such as tune and projection, can be similarly addressed: tune assigns CB to right siblings of a pitch accent, while projection non-deterministically extends an assignment leftward within so-constrained limits.

4.3 Case studies

In this section we apply our formalization to various examples, both illustrating how the theory of §4.1 works out and how it relates to other frameworks.

We start with a few simple examples. Throughout this section we present the inferences from the principles in a tabular fashion, with the **T/F** column showing the inferred CB/NB-ness of each referent.

(16) (What did you do?)
　　　I gave Kathy a BOOK.

For (16) we have the following inferences.

Word	Ctxt	SO	Tune	Det	Proj	T/F
I	CB	T	T	T	T	CB
gave	T	T	T	T	NB	NB
Kathy	T	T	T	T	NB	NB
book	T	T	NB	NB	T	NB

(17) presents a variation on (16), with a topicalized PATIENT. The inferences are given in the table.

(17) (What did you do with the book?)
　　　The book, I gave to KATHY.

Word	Ctxt	SO	Tune	Det	Proj	T/F
book	CB	CB	T	CB	T	CB
I	CB	T	T	T	T	CB
gave	T	T	T	T	NB	NB
Kathy	T	T	NB	T	T	NB

Now, consider again (8a,b), repeated as (18a,b).

(18) (On Tuesday, what flight did John take?)
　　　a. On Tuesday, John flew from London to PARIS
　　　b. # On Tuesday, John flew to PARIS from London.

For (18a) we get the following inferences from the different principles, and the context.

Word	Ctxt	SO	Tune	Det	Proj	T/F
Tuesday	CB	CB	T	T	T	CB
John	CB	T	T	T	T	CB
flew	CB	T	T	T	T	CB
London	T	T	T	T	NB	NB
Paris	T	T	NB	T	T	NB

(18a) is similar to (16): The topicalization of "on Tuesday" makes it CB, whereas the pitch accent on "Paris" indicates it is NB. In the end, projection makes "from London" CB.

For (18b) we get a different analysis, correctly inferring it is dispreferred.

Word	Ctxt	SO	Tune	Det	Proj	T/F
Tuesday	CB	CB	⊤	⊤	⊤	CB
John	CB	⊤	⊤	⊤	⊤	CB
flew	CB	⊤	⊤	⊤	⊤	CB
Paris	⊤	CB	NB	⊤	⊤	⊥
London	⊤	⊤	CB	⊤	⊤	CB

Due to the pitch accent on "Paris", we infer that "Paris" is NB and that "from London" (as its right-adjacent sister) is CB. However from SO we also infer that "Paris" is CB, resulting in a conflict, providing one ground to rule out the example. Another ground would result from further discourse interpretation: "London" cannot be interpreted as CB, as it has not been activated in the context.

To illustrate embedded foci, consider (19).

(19)   (Which teacher did you give what book?)
       I gave the book ON SYNTAX to the lecturer OF EN-
       GLISH.

Word	Ctxt	SO	Tune	Det	Proj	T/F
I	CB	⊤	⊤	⊤	CB	CB
gave	CB	⊤	⊤	⊤	CB	CB
book	CB	⊤	⊤	CB	⊤	CB
syntax	⊤	⊤	NB	⊤	⊤	NB
teacher	CB	⊤	⊤	⊤	⊤	CB
English	⊤	⊤	NB	⊤	⊤	NB

The pitch accents on "syntax" and "English" establish them as NB, though not determining "teacher" as CB since "teacher" is not a sibling of "syntax". Using projection we can confirm "I" and "gave" being CB, given that "the book" is CB on account of the definite determiner.

Information packaging (Vallduví, 1990) is unable to establish a topic and focus for (19), due to the embedding coupled with discontinuity. Using our recursive procedure, we have no such problems, arriving at a focus being constituted by "syntax" and "English".

Finally, we turn to the Dutch examples. We only examine the variations in (5), repeated here as (20); the all-focus case in (4) is trivial, projecting NB leftwards from the sentence-final pitch accent.

(20)   a. (Who read "Moby Dick"?)
          JAN_NB las_CB "Moby Dick"_CB.

       b. (Who read "Moby Dick"?)
          "Moby Dick"_CB las_CB JAN_NB.

       c. (What did John read?)
          Jan_CB las_CB "MOBY DICK"_NB.

       d. (What did John read?)
          "MOBY DICK"_NB las_CB Jan_CB.

The analysis of (20a) is as follows. Observe that the dependents are ordered canonically, hence the SO principle yields only ⊤.

Word	Ctxt	SO	Tune	Det	Proj	T/F
Jan	⊤	⊤	NB	⊤	⊤	NB
las	CB	⊤	⊤	⊤	⊤	CB
Moby Dick	CB	⊤	⊤	⊤	⊤	CB

The analysis of (20b) differs from the one for (20a) because of the order variation. The SO principle now assigns CB to "Moby Dick", while the pitch accent on "Jan" again makes it NB.

Word	Ctxt	SO	Tune	Det	Proj	T/F
Moby Dick	CB	CB	⊤	⊤	⊤	CB
las	CB	⊤	⊤	⊤	⊤	CB
Jan	⊤	⊤	NB	⊤	⊤	NB

For the analysis of (20c) given below, observe that in the given context it is the contextual activation of "Jan" and "las" that prevent the projection principle to assign NB to the referents leftwards of "Moby Dick".

Word	Ctxt	SO	Tune	Det	Proj	T/F
Jan	CB	⊤	⊤	⊤	⊤	CB
las	CB	⊤	⊤	⊤	⊤	CB
Moby Dick	⊤	⊤	NB	⊤	⊤	NB

Finally, consider (20d). Our principles predict that a referent with a pitch accent is NB, while a referent violating SO is CB – both cannot be simultaneously the case. Thus, in general a dependent that appears sentence-initial, and which receives pitch accent, must fill a semantic role that is leftmost in its head's SO. In a declarative sentence in active voice this typically is the ACTOR. This is why (20d) is ruled out, as the analysis below shows.

Word	Ctxt	SO	Tune	Det	Proj	T/F
Moby Dick	⊤	CB	NB	⊤	⊤	⊥
las	CB	⊤	⊤	⊤	⊤	CB
Jan	CB	⊤	⊤	⊤	⊤	CB

Because (Hoffman, 1995) or (Hajičová et al., 1995) provide no account in which word order and tune are integrated, it is difficult to see how they would deal with the examples above. Using different lexical entries to deal with the word order variations in (20), (Steedman, 2000) is in principle able to deal with these examples. However, CCG lacks the mechanisms to extend the account to the degree of word order freedom found e.g. in German – whereas TDG is able to do so (Duchier and Debusmann, 2001).

4.4 Final remarks

The lattice-based model presented in this paper is of course only an idealization. A more realistic and robust model will need to appeal to preferences. However, considerable mileage can be derived from slightly more elaborate lattices that capture essential aspects of preference models.

5 Conclusions

In this paper, we presented an extension to Topological Dependency Grammar to address the derivation of an expression's information structure. We indentified a number of principles which on the basis of structural indications of informativity contribute to the determination of CB/NB-ness. Our contribution is two-fold: first, our principles derive evidence of CB/NB-ness in the 4-valued Boolean lattice, thus supporting both underspecification by lattice top ⊤ and easy combination by lattice meet ⊓; second we have shown that our formulation naturally fits in the concurrent constraint approach of TDG. As a consequence, we have access to practically efficient constraint-based parsers, and we take advantage of the fact that multiple sources of structural indications can simultaneously influence the realization of information structure. In this, we reach beyond existing approaches such as (Steedman, 2000), (Hoffman, 1995), or (Hajičová et al., 1995).

The approach is conceptually related to (Kruijff, 2001), who presents a framework in which different types of structural indications can interact. However, Kruijff's framework does not come with an efficient implementation, and is formally more intricate than the constraint-based approach we present here.

One topic for further research is how to derive a logical representation from the analysis we now obtain, similar to (Kruijff, 2001; Copestake et al., 1999) or (Baldridge and Kruijff, 2002). Having a logical representation would provide a convenient bridge to discourse interpretation.

References

Jason Baldridge and Geert-Jan Kruijff. 2002. Coupling CCG and hybrid logic dependency semantics. In *Proceedings ACL'02*, pages 319–326, Philadelphia, Pennsylvania.

Ann Copestake, Dan Flickinger, and Ivan A. Sag. 1999. Minimal recursion semantics. an introduction. Unpublished Manuscript. CSLI/Stanford University.

Denys Duchier and Ralph Debusmann. 2001. Topological dependency trees: A constraint-based account of linear precedence. In *Proceedings ACL'01*, Toulouse, France.

Denys Duchier. 2001. Lexicalized syntax and topology for non-projective dependency grammar. In *MOL 8 Proceedings*.

Eva Hajičová, Barbara H. Partee, and Petr Sgall. 1998. *Topic-Focus Articulation, Tripartite Structures, and Semantic Context*. Kluwer Academic Publishers.

Eva Hajičová, Hana Skoumalová, and Petr Sgall. 1995. An automatic procedure for topic-focus identification. *Computational Linguistics*, 21(1):81–94, March.

Beryl Hoffman. 1995. Integrating "free" word order syntax and information structure. In *Proceedings EACL'95*, Dublin, March.

Ivana Kruijff-Korbayová, Geert-Jan M. Kruijff, and John Bateman. 2002. Generation of contextually appropriate word order. In Kees van Deemter and Roger Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 193–221. CSLI Publications, Stanford CA.

Ivana Kruijff-Korbayová, Stina Ericsson, Kepa-Joseba Rodríguez, and Elena Karagjsova. 2003. Producing contextually appropriate intonation in an information-states based dialogue system. In *Proceedings EACL'03*, Budapest, Hungary.

Geert-Jan M. Kruijff. 2001. *A Categorial-Modal Logical Architecture of Informativity: Dependency Grammar Logic & Information Structure*. Ph.D. thesis, Charles University, Prague, Czech Republic.

Daniela Kurz, Wojciech Skut, and Hans Uszkoreit. 2000. German factors constraining word order variation. In *Thirteenth Annual Conference on Human Sentence Processing CUNY 2000*, La Jolla, California.

Scott Prevost and Mark Steedman. 1994. Specifying intonation from context for speech synthesis. *Speech Communication*, 15(1–2):139–153.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company.

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

Enric Vallduví and Elisabet Engdahl. 1996. The linguistic realization of information packaging. *Linguistics*, 34:459–519.

Enric Vallduví. 1990. *The Informational Component*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.