Information sharing

Book editors

October 20, 2000

**CENTER FOR THE STUDY
OF LANGUAGE
AND INFORMATION**

Contents

1

Generation of Contextually Appropriate Word Order

1.1 Introduction

In any natural language communication, the content that is being communicated is by nature multi-dimensional (in the sense that conceptual structures are not linear). However, since each participant can only utter the elements in a linear order, linguistic structures of all degrees of complexity need to be projected into one dimension, and the individual elements must be ordered sequentially according to certain rules. The rules and schemes of linear ordering can be studied at different strata of a given language. What we concentrate on in this paper are the rules and schemes pertaining to the linear ordering of elements constituting a clause. The linear ordering of clause elements is usually referred to as **word order** (WO), even though the elements are not only individual words. We consider both the mutual ordering of phrases and groups within a clause, and their ordering with respect to the main verb.

The problem of proper WO treatment is a very important and challenging one in **natural language generation** (NLG), especially when the generation of connected, natural text (spoken or written) is concerned. Without a correct WO any generated text is much less fluent than it should be, and wrong WO can even lead to misinterpretation. Well known examples for English include interpretations that assign differing preferred quantifier scopes depending on the order in which the linguistic expressions introducing those quantifiers appear in the surface string; others include controlling preferred sites of prepositional phrase or relative clause attachment by careful ordering. These problems, significant though they are for English, become substantially more important when we consider languages with a higher degree of **word order freedom**, like the Slavonic languages. The issue of WO control for such languages must be addressed in NLG because the various WO variants of a sentence, although *grammatically* well-formed, do not necessarily have the same meaning and are generally not interchangeable in a given context. The following example illustrates this with two versions of a short discourse in Czech: The WOs throughout (2) are bizarre, and do not correspond to the intended meaning. With the exclusion of very particular intonation patterning, this text is even unacceptable and incoherent.[1]

(1) a. Otevřete nový obrázek.
open$_{i2pl}$ new drawing$_{acc}$
"Open a new drawing."

b. Vložte čáru a oblouk.
insert$_{i2pl}$ line$_{acc}$ and arc$_{acc}$
"Insert a line and an arc."

c. Nyní obrázek uložte.
now drawing$_{acc}$ save$_{i2pl}$
"Now save the drawing."

(2) a. Nový obrázek otevřete.
new drawing$_{acc}$ open$_{i2pl}$
"Open the new drawing."

b. Čáru a oblouk vložte.
line$_{acc}$ and arc$_{acc}$ insert$_{i2pl}$
"Insert the line and the arc."

c. Nyní uložte obrázek.
now save$_{i2pl}$ drawing$_{acc}$
"Now save a drawing."

The fact that WO is relatively free in the Slavonic languages, but not arbitrary with respect to a given communicative intention and the current discourse context, brings about two important issues for the process of automatic generation of continuous texts:

- **coverage**: the generator (the grammar) needs to be able to generate all the grammatically well-formed WO variants in a given language
- **control**: the generator must choose the most appropriate grammatically well-formed WO with respect to the communicative intention and the context

From a cross-linguistic perspective, or from the viewpoint of building a multilingual NLG system, an additional issue comes in:

- **flexibility**: WO in different languages should be described in uniform terms to enable cross-linguistic generalizations; furthermore, it is desirable to employ a uniform mechanism handling WO, which can be adjusted to a particular language by setting a few parameters, or by providing language-specific preferences.

Our discussion in this paper is anchored within the context of developing a multilingual system for generating software instruction texts in

[1]Throughout the paper, we assume a neutral intonation pattern in all examples in the Slavonic languages, i.e., the intonation centre coinciding with the last element of a clause. In some cases, the corresponding English sentence may require a marked placement of the intonation center.

Czech, Bulgarian and Russian. We concentrate on the factors determining WO in these languages, and the corresponding issues of WO control in multilingual generation. Nevertheless, the issues we address and the WO generation approach we propose are applicable also for other languages and are relevant in any register.

Outline of the paper In Section 1.2 we first demonstrate the degree of WO freedom with which one is confronted in Czech, Bulgarian and Russian, by presenting several examples from software instruction texts.

Section 1.3 presents the control issue in more detail. We first give examples supporting the claim that WO in the three Slavonic languages is not arbitrary. Then, we concentrate on the factors restricting WO in any particular case, namely global text organization, information structure, and syntactic structure.

In Section 1.4 we discuss thematic and information structures as a source of WO control. We combine the insights concerning determination of WO in two functionally-oriented approaches: Halliday's (1985) Systemic Functional Grammar framework and the framework of Functional Generative Description (Sgall *et al.* 1986).

In Section 1.5 we present our approach to generating WO; this approach is sufficiently flexible as well as sufficiently constrained, so as to be suitable for languages with a varying degrees of WO freedom. We demonstrate the application of our approach on Czech in Section 1.5.2.

Section 1.6 closes the paper with a brief discussion of the overall appropriateness of the proposed approach.

1.2 Word Order Variation

For sentences with a moderate number of syntactic groups there are usually quite a few WO variants in the Slavonic languages that are grammatically well-formed but may differ in the contextual conditions of their use (see Section 1.3). Let us consider the following sentences containing one main verb and two or more nominal groups or prepositional phrases appearing either before or after the verb in different orders.[2]

[2]Note some of the differences between the languages: Czech and Russian have richer morphology, and distinguish seven and six cases, respectively; Bulgarian does not differentiate between cases. In the discussed examples, instrument is expressed by instrumental case in Czech and Russian, but it is expressed using the preposition 'с' in Bulgarian. Another difference concerns articles: whereas in Czech and Russian there is no definite or indefinite article (although pronouns can be used as determiners), Bulgarian has a definite suffix, which appears as '-та' in 'командата' in the discussed examples. In spite of these differences, we present just one common gloss in the examples below, for simplicity.

(3) **Cz** Otevřete soubor příkazem `Open`.
Ru Откройте файл коммандой `Open`.
Bg Отворете файла с командата `Open`.
gloss open$_{i2pl}$ file$_{acc}$ command$_{ins}$ `Open`
En Open a file with the `Open` command.

(4) **Cz** Soubor otevřete příkazem `Open`.
Ru Файл откройте коммандой `Open`.
Bg Файла отворете с командата `Open`.
gloss file$_{acc}$ open$_{i2pl}$ command$_{ins}$ `Open`
En Open the file with the `Open` command.

(5) **Cz** Příkazem `Open` otevřete soubor.
Ru Коммандой `Open` откройте файл.
Bg С командата `Open` отворете файла.
gloss command$_{ins}$ `Open` open$_{i2pl}$ file$_{acc}$
En With the `Open` command, open a file.

(6) **Cz** Příkazem `Open` soubor otevřete.
Ru ? Коммандой `Open` файл откройте.
Bg * С командата `Open` файла отворете.
gloss command$_{ins}$ `Open` file$_{acc}$ open$_{i2pl}$
En With the `Open` command, open the file.

(7) **Cz** Otevřete příkazem `Open` soubor.
Ru ? Откройте коммандой `Open` файл.
Bg * Отворете с командата `Open` файла.
gloss open$_{i2pl}$ command$_{ins}$ `Open` file$_{acc}$
En Open the file with the `Open` command.

(8) **Cz** Soubor příkazem `Open` otevřete.
Ru * Файл коммандой `Open` откройте.
Bg * Файла с командата `Open` отворете.
gloss File$_{acc}$ command$_{ins}$ `Open` open$_{i2pl}$
En Open the file with the `Open` command.

In all three languages, (3,4,5) are grammatical. (6) and (7) are grammatical in Czech, but unnatural in Russian, and ungrammatical in Bulgarian. (8) is grammatical in Czech, but ungrammatical in Russian and Bulgarian. We encountered differences in grammaticality judgements on WO variants in Bulgarian, Czech and Russian also in some other cases (cf. (Adonova *et al.* 1999) for more discussion). However, a more thorough study would be needed to confirm and generalize such observations.

The instructions in (3-8) are expressed by (polite) imperatives (in second person plural). In Czech, it is also possible to formulate instructions in indicative mood with active voice. The Actor in the Subject position can be realized by a zero-pronoun,[3] as in (9). The range of possible WO

[3]In instruction texts in Czech, first person plural is usually used in indicative mood, less often second person plural. Second person singular can be used only in informal style.

variation in the indicative mood is the same as in the imperative mood, i.e., all six permutations are grammatically well-formed.

(9) **Cz** Otevřeme soubor příkazem `Open`.
open$_{1pl}$ file$_{acc}$ command$_{ins}$ `Open`
"One opens a file with the `Open` command."

When the Subject is non-pronominal, it undergoes the same ordering mechanisms as other elements in the clause. The Czech sentences in (10,11) thus have twenty four and six WO permutations, respectively, all of which are grammatically well-formed.

(10) **Cz** Uživatel otevírá soubor příkazem `Open`.
user$_{nom}$ open$_{3sg}$ file$_{acc}$ command$_{ins}$ `Open`
"The user opens a file with the `Open` command."

(11) **Cz** Příkaz `Open` otevírá soubor.
command$_{nom}$ `Open` open$_{3sg}$ file$_{acc}$
"The `Open` command opens a file."

Furthermore, all three languages can express instructions by sentences in indicative mood in passive voice, using a reflexive passive construction, as in (12) where we show one WO variant for illustration.

(12) **Cz** Příkazem `Open` se otevírá soubor.
command$_{ins}$ `Open` *refl* open$_{3sg}$ file$_{nom}$
Bg С командата `Open` се отваря файл.
with command$_{def}$ `Open` *refl* open$_{3sg}$ file
Ru Коммандой `Open` откроется файл.
command$_{ins}$ `Open` open$_{refl3sg}$ file$_{nom}$
"With the `Open` command, a file is opened."

The use of the reflexive passive brings about the problem of **clitic placement**, in particular the placement of reflexive particles in Czech and Bulgarian.[4] This is an example of grammatically constrained ordering. In Bulgarian, the particle "se" is placed either after or before the main finite verb form in a clause, depending on whether the verb is clause-initial or not, respectively:

(13) **Bg** Отваря се файл.
open$_{3sg}$ *refl* file
"A file opens."

(14) **Bg** Файл се отваря.
file *refl* open$_{3sg}$
"A file opens."

In Czech the situation is more complicated: Czech clitics have to be placed in the so-called **Wackernagel position**, characterized roughly

[4]In Russian, reflexive forms are created with the '-sja' suffix attached to the main verb, and therefore no particular ordering issue arises.

as the position between the first and the second element in a clause.[5] (15-19) show the other grammatically well-formed ordering alternatives corresponding to the Czech sentence in (12).

(15) **Cz** Příkazem `Open` se soubor otevírá.
Command$_{ins}$ `Open` *refl* file$_{nom}$ open$_{3sg}$
"With the `Open` command, the file is opened."

(16) **Cz** Soubor se otevírá příkazem `Open`.
file$_{nom}$ *refl* open$_{3sg}$ command$_{ins}$ `Open`
"The file is opened with the `Open` command."

(17) **Cz** Soubor se příkazem `Open` otevírá.
file$_{nom}$ *refl* command$_{ins}$ `Open` open$_{3sg}$
"The file is opened with the `Open` command."

(18) **Cz** Otevírá se soubor příkazem `Open`.
opens *refl* file$_{nom}$ command$_{ins}$ `Open`
"A file is opened with the `Open` command."

(19) **Cz** Otevírá se příkazem `Open` soubor.
opens *refl* command$_{ins}$ `Open` file$_{nom}$
"A file is opened with the `Open` command."

If there are multiple clitics in a single clause in Czech, then they have to be grouped in a cluster. Their mutual ordering depends on their functions and is constrained by the grammar.

The important conclusion drawn from these comparisons of WO freedom is that in developing a uniform mechanism for handling WO in our multilingual NLG system, we need to allow for a sufficient degree of flexibility such that each grammar can set whatever grammatical constraints are needed in the individual language, while at the same time, leaving grammatically unconstrained ordering to be decided by other appropriate strategies (cf. Section 1.5).

1.3 Context and WO

Various factors have been discerned in the language system in general which play an important role in expressing a given content in a specific linear form. The inventory of these factors contains at least the following: *information structure, syntactic structure, intonation, rhythm* and *style*. These factors are very general, and can therefore be considered

[5]Naturally, this leaves to be defined what "element" means. It is easy to show that "first element" does not equate to "first constituent" in any simple sense, since the element can be of arbitrary complexity. It is also possible to find examples where multiple elements precede the clitic cluster; however, these cases seem to be restricted to cases where all those pre-clitic elements are strongly backgrounded (Oliva and Avgustinova 1995).

language universals, certainly within the family of Indo-European languages. However, the individual factors may have different importance for the linear ordering in a given language.

For instance, English is an example of a language where WO is strongly constrained by syntactic structure. In such languages, differences in information structure are often reflected by varying the intonation pattern or by the choice of a particular type of grammatical construction. This latter includes, e.g., the use of a definite nominal group to refer anaphorically, and the use of an indefinite nominal group to introduce new entities. In languages with rich cases systems and no definite or indefinite article, such as Czech and Russian, the same effects are achieved by varying WO in accordance to information structure. In contrast, Bulgarian is closer to English in that its WO is more rigid (cf. examples (1) and (2) in Section 1.1).

In order to illustrate the general point that WO selections are related to information structure, let us overview appropriateness judgements for the Czech versions of examples (3-8) presented in Section 1.2, repeated below for convenience. Similar judgments hold in all three languages, although the three languages differ, as we noted in Section 1.2 above, in grammaticality judgments of the sentences.

(20) **Cz** Otevřete příkazem `Open` soubor.
open$_{i2pl}$ command$_{ins}$ `Open` file$_{acc}$
"Open a file with the `Open` command."

The ordering in (20) is neutral with respect to context (cf. 'systemic ordering' in Section 1.4.1, 11). It can be used "out of the blue", or in a context which can be approximated by the question *What should we do?*.[6] Alternatively, (20) could be used in a context characterized by the question *What should we open by the* `Open` *command?*, when the `Open` command is not being contrasted with some other entity.

(21) **Cz** Otevřete soubor příkazem `Open`.
open$_{i2pl}$ file$_{acc}$ command$_{ins}$ `Open`
"Open the file with the `Open` command."

(22) **Cz** Soubor otevřete příkazem `Open`.
file$_{acc}$ open$_{i2pl}$ command$_{ins}$ `Open`
"Open the file with the `Open` command."

Both (21) and (22) are appropriate when some file is salient, for instance when the user is working with a file. That is why we put the

[6]We use questions to indicate which contexts would be appropriate for uttering sentences with particular WO variants. Such question-answer pairs should be conceived of in terms of the **question test** (Sgall *et al.* 1986) rather than as constituting natural dialogue exchanges.

definite article into the English translation. The action of opening can, but does not have to, be salient, too. The contexts in which (21) and (22) can be appropriately used can be characterized by the questions *What should we do with the file?* or *How should we open the file?*. Unlike (21), (22) can also be used if 'soubor' (file) is contrasted with another entity.

(23) **Cz** Příkazem `Open` otevřete soubor.
command$_{ins}$ `Open` open$_{i2pl}$ file$_{acc}$
"With the `Open` command, open a file."

(24) **Cz** Příkazem `Open` soubor otevřete.
command$_{ins}$ `Open` file$_{acc}$ open$_{i2pl}$
"With the `Open` command, open the file."

The contexts in which (23) can be used are characterized by *What should we do with the* `Open` *command?*. While (23) does not refer to a specific file, in (24) a salient file is presumed. (24) is appropriate in contexts characterized by *What should we do to the file with the* `Open` *command?*.

It is also possible to use (23) in a context characterized by *What should we do?*, and (24) in a context characterized by *What should we do to the file?*, if it is presumed that we are talking about using various commands (or various means or instruments) to do various things. In the latter type of context, the `Open` command does not have to be salient.

(25) **Cz** Soubor příkazem `Open` otevřete.
file$_{acc}$ command$_{ins}$ `Open` open$_{i2pl}$
"Open the file with the `Open` command."

Example (25) is appropriate when both a file and the `Open` command are salient. The contexts in which (25) can be appropriately used can be characterized by *What should we do to the file with the* `Open` *command?*.

Analogous judgements concerning contextual appropriateness apply to the variants of these instructions in indicative mood, both in active voice and in the reflexive passive, cf. (9-19) in Section 1.2. The orders in which the verb is first do not presume the salience of either a file or a command. The orders in which 'file' precedes the verb appear to presume the salience of a file, the order in which 'command' precedes the verb appear to presume the salience of a command. When both 'file' and 'command' precede the verb, the salience of both a file and a command appears to be presumed.

These judgements show that differences in WO very often correspond to differences in informative status of the entities and processes that the text concerns: in particular, whether they are already familiar or not, and whether they are assumed to be salient. Note that in English, a definite vs. indefinite nominal expression, i.e. 'a|the file', is used in the individual WO variants.

The extent to which WO reflects information structure is constrained by the degree of WO freedom in a given language. In Czech and Russian, as expected, WO reflects information structure in most cases (cf. (Adonova *et al.* 1999) for exceptions observed in Russian). Bulgarian resembles English in its preference for a single element preceding the main verb of a clause. On the other hand, the ordering of the elements following the verb has a higher degree of freedom, at least in some configurations. We hypothesize that the Bulgarian WO follows information structure in this case.

To summarize so far: Since sentences which differ only in WO (and not in syntactic realizations of constituents) are not freely interchangable in a given context, we have to be able to generate contextually appropriate WO. In order to achieve this, we need to be able to capture not only the structural restrictions specific for the individual languages, but also the information status of the entities being referred to.

1.4 Sources for Contextual Control of Word Order

In the Slavonic linguistic tradition, **information structure** has always been considered the main factor determining the linear ordering within a sentence (cf. the works of Mathesius (1939), the Prague School (Sgall *et al.* 1973; Sgall *et al.* 1986; Hajičová and Sgall 1987; Hajičová 1993); or Firbas (1992)). Although "information structure" can be used as a general term for various notions employed in contemporary theories of the syntax-semantics interface—notions that reflect how the conveyed content is distributed over a sentence, and how it is thereby structured or "perspectivized"[7]—we also need it to become sufficiently focused so as to support an implemented generation system.

This task is complicated by the fact that many terminological frameworks have been proposed in the literature, but it is in general unclear to what extent these varying accounts are mutually compatible. Comparison and use is further complicated by the fact that similar terms are also often used (not always knowingly) to refer to quite different theoretical entities. For example, in every approach to information structure, the clause is considered to consist of (at least) two parts: The often used dichotomies are, e.g. *Theme-Rheme* (Sgall *et al.* 1973; Halliday 1985; Daneč 1985; Firbas 1992; Steedman to appear), *Topic-Focus* (Sgall *et al.*

[7]The term information structure has been coined by Halliday (1967), and is explicitly used for example by Steedman (1994; 1996; 2000; to appear). Consider also Chafe's *information packaging* as used by Vallduví (1992), *topic-focus articulation* within the Modern Prague School (Sgall *et al.* 1986), or Firbas' (1992) *functional sentence perspective*.

1986), *Background-Focus* (Krifka 1992), *Ground-Focus* (Vallduví 1992).[8] But there are substantial differences between the individual usage of particular researchers.

In order to provide constraints for WO decisions within our generation architecture, we require mechanisms through which particular patterns of information structuring can appropriately influence the choice among the WO variants available. These patterns are provided by our text planning component. We have found two complementary approaches to the relationship between aspects of information structuring and WO to be ripe for application in the generation of extended texts. In this section, we briefly introduce them and show their application.

In order to clarify the 'complementary' nature of the approaches that we have adopted, it is necessary first to distinguish between two dimensions of organization that are often confused or whose difference is contested: in his Systemic Functional Grammar (SFG), Halliday (1970;1985) distinguishes between the **thematic structure** of a clause and its **information structure**.

The thematic structure and information structure of a clause are closely related but not the same. Whereas the **Theme** "is the starting point for the message, it is the ground from which the clause is taking off" (Halliday 1985: p.38), the information structure concerns the distinction between the **Given** as "what is presented as being already known to the listener" (Halliday 1985: p.59), and the **New** as "what the listener is being invited to attend to as new, or unexpected, or important" (*ibid*).

1.4.1 Information Structure and Ordering

In Halliday's original approach, the root assumption for English is that ordering, when not grammatically constrained, is iconic with respect to "newsworthiness"—which is part of the information structure and hence concerns Given/New. So under the scale from Given to New, the "newer" elements would come towards the end of the information unit; the "newest" element would be the bearer of nuclear stress. This approach relies on the possibility of giving a complete ordering of all elements in the context with respect to their salience or newsworthiness. But, to our knowledge, this approach has not yet been applied in a complete NLG system. Although Steedman's (1996) elaboration of Halliday's ideas concerning information structure, for example, has been applied to generation of spoken English (Prevost and Steedman 1994; Prevost 1995), the main concern of that work was to assign contextually

[8] Within the scope and aim of the present paper, we have to refrain from a more in-depth discussion of the various approaches to information structure (see, e.g., (Hajičová and Kruijff-Korbayová 1999)).

appropriate intonation to answers in a dialogue system. The ordering of clause elements was fully determined by syntactic structure.

The notion of ordering by newsworthiness in Halliday's approach is parallel to the notion of **communicative dynamism** (CD) introduced in the early works of Firbas (for a recent formulation see (Firbas 1992)) and used also within the Functional Generative Description (FGD, Sgall *et al.* 1986). In this view, too, the prototypical ordering of clause elements from left to right respects newsworthiness: in prototypical cases, WO corresponds to CD. However, textually motivated thematization or grammatical constraints may force WO to diverge from CD.

The FGD approach differs from Halliday's in that, in addition to CD, it works with a default (canonical) ordering, called **systemic ordering** (SO). SO is a *language specific* ordering of complements, i.e. "arguments" and "adjuncts", of verbs, nouns, adjectives or adverbs which corresponds to the canonical ("neutral") WO. SO is considered constant within a given language. SO has been studied in most detail for Czech (Sgall *et al.* 1980; Sgall *et al.* 1986; Hajičová and Sgall 1987; Petkevič 1995); for SO in English see (Hajičová *et al.* 1995), for SO in German see (Sgall *et al.* 1995). Similar ideas for Japanese have been developed for instance by Kuno (1972). The SOs for Russian and Bulgarian have not yet been studied in general. We follow Sgall *et al.* (1986) in considering the SOs for the main types of complementations in Russian and Bulgarian to be similar to the Czech one, though there can be slight differences (cf. the observations reported in (Adonova *et al.* 1999)).

For our current purpose we concentrate on the SO for a subset of the complement types that are discerned in FGD. The initial versions of the SO we work are shown below:[9].

(26) SO for Czech, Russian, Bulgarian:
Actor < TemporalLocative < Purpose < SpaceLocative < Means < Addressee < Patient < Source < Destination

(27) SO for English:
Actor < Addressee < Patient < SpaceLocative < TemporalLocative < Means < Source < Destination < Purpose-dependent

The SO for the Slavonic languages is based on that for Czech presented by Sgall *et al.* (1986); the only difference is that we have placed Patient

[9]The labels we use for the various types of complements are a mixture of FGD and SFG terminology. We unfortunately do not have the space to describe the complement types here or to present out a detailed comparison between the complements types distinguished in FGD and the Participants and Circumstances distinguished in SFG (cf. (Panevová 1974; Halliday 1985; Sgall *et al.* 1986; Petkevič 1995; Kruijff forthcoming). For an illustration, see examples (28-42).

before Source ('from where'). The SO for English combines the suggestions made by Sgall *et al.* (1986) and the ordering defaults of the NIGEL grammar of English (cf. Section 1.5.2).

Informational prominence (in the negative and positive sense) is established through deviation of CD from the SO background. This leads us to the distinction FGD makes between **contextually bound** (CB) and **contextually non-bound** (NB) items in a sentence (Sgall *et al.* 1986): A CB item is assumed to convey some content that bears a contextual relationship to the discourse context. Such an item may refer to an entity already explicitly referred to in the discourse, or an "implicitly evoked" entity (Hajičová 1993). At each level of syntactic structure, CB items are ranked lower than NB items in the CD ordering. The motivation behind and the meaning of the CB/NB distinction in FGD corresponds to those underlying the Given/New dichotomy in SFG.

Contextual boundness is a local feature, i.e., a complex CB item can contain "locally" NB items, and vice versa (Petkevič 1987). For instance, constituents within clauses in a complex sentence are discerned as locally CB vs. NB. Also within a complex nominal group which is, e.g., NB as a whole, some parts can be CB (a straightforward example of this kind of nominal group is an NB noun modified by a CB possessive pronoun). In (32) below, both 'soubor' (file) and 'editor' (editor) are locally CB.

The following examples illustrate the relation between WO, CD, SO and CB/NB. The types of arguments and adjuncts, as well as the CB/NB assignment, are indicated in subscripts. The WO in the Czech versions respects CD, while in the English versions the WO diverge from CD due to grammatical constraints. Note also that there are differences in ordering due to the difference in SO between Czech and English; e.g., Means precedes Patient in the SO in Czech, but follows it in English.

(28) Přeneseme myší ikonu z
drag$_{1pl}$ mouse$_{ins}{}^{Means-NB}$ icon$_{acc}{}^{Patient-NB}$ from
plochy do editoru.
surface$_{gen}{}^{Source-NB}$ into editor$_{gen}{}^{Destin-NB}$
We drag the icon with the mouse from the desktop to the editor.

(29) Ikonu přeneseme myší z
icon$_{acc}{}^{Patient-CB}$ drag$_{1pl}$ mouse$_{ins}{}^{Means-NB}$ from
plochy do editoru.
desktop$_{gen}{}^{Source-NB}$ into editor$_{gen}{}^{Destin-NB}$
The icon we drag with the mouse from the desktop to the editor.

(30) Myší z plochy do editoru
mouse$_{ins}{}^{Means-CB}$ from desktop$_{gen}{}^{Source-CB}$ into editor$_{gen}{}^{Destin-CB}$
přeneseme ikonu.
drag$_{1pl}$ icon$_{acc}{}^{Patient-NB}$
With the mouse we drag the icon from the desktop to the editor.

(31) Ikonu myší do editoru
icon$_{acc}$$^{Patient-CB}$ mouse$_{ins}$$^{Means-CB}$ into editor$_{gen}$$^{Destin-CB}$
přeneseme z plochy.
drag$_{1pl}$ from desktop$_{gen}$$^{Source-NB}$

The icon we drag with the mouse to the editor from the desktop.

(32) Pro otevření souboru ikonu souboru přeneseme
(For opening file$_{gen}$)$^{Purpose-CB}$ (icon$_{acc}$ file$_{gen}$)$^{Patient-CB}$ drag$_{1pl}$
myší z pracovní plochy do
mouse$_{ins}$$^{Means-NB}$ (from work$_{adj}$ desktop$_{gen}$)$^{Source-NB}$ (into
okna editoru.
window$_{gen}$ editor$_{gen}$)$^{Destin-NB}$

In order to open the file, we drag the file's icon with the mouse from the desktop to the editor window.

From a generation point of view, the ordering variants illustrated above are fully determined by the CB/NB assignment, the SO and the principles of ordering CB elements before NB ones and placing the verb between the CB and the NB elements. (28) shows the basic case where the only CB element is the Actor (realized by the implicit (dropped) Subject in Czech), all other elements are NB, and their ordering follows the SO. Although WO is generally much more restricted in English than in Czech, it is interesting to observe that there are differences as well as parallels in ordering between the Czech and English sentences.

In (29), which should be seen as an answer to *What do we do with the icon using the mouse?*, both Czech and English allow ordering the CB element 'icon' before the verb. In (30), which should be seen as an answer to *What do we drag from the desktop to the editor using the mouse?*, and in (31), which should be seen as an answer to the question *Where from do we drag the icon to the editor using the mouse?*, Czech but not English allows ordering all the CB elements before the verb.

The final example (32) is an extension of (29), with an additional CB element, namely Purpose, and with more complex constituents. The latter are enclosed in round brakets for clarity. For simplicity, we do not mark local CB/NB, because it is not relevant for the discussion in this paper. We can again observe the CB/NB motivated WO in Czech, and a difference in ordering between Czech and English.

Note that the CB/NB distinction enables us to explain and motivate the difference in ordering between the English version of (29) and (31), following the discussion of similar examples in (Sgall *et al.* 1986):

(33) a. They flew from Chicago to Boston.
b. They flew to Boston from Chicago.

(34) a. They went by car to a river.
b. They went to a river by car.

In (33a) and (34a), the ordering of the circumstantial modifiers after the verb is in accordance with the SO (cf. p. 11): Source and Means precede Destination, respectively, while in (33b) and (34b) this is not the case. Sgall *et al.* attribute the difference in WO to information structure: both circumstantial modifiers in (33a) and in (34a) are NB, and therefore their order respects the SO. On the other hand, the Source in (33b) and the Means in (34b) are CB, and therefore they are ordered before the Destination which remains NB in both cases (under neutral intonation).

Likewise, the CB/NB distinction enables us to explain and motivate the difference in ordering between (35) and (36).

(35) He picked up the pen.

(36) He picked the pen up.

Such **particle movement** constructions are discussed, for example by Gries (1999). Within a cognitive approach, he attributes the difference in ordering to the degree of effort needed to "process" the direct object. Our explanation in terms of CB/NB is as follows: We treat the particle as expressing the Destination (which is ranked higher on the SO scale than the Patient, cf. p. 11). It can be either CB or NB.[10] (35) is then an ordering corresponding to 'up' being CB and 'pen' being NB, e.g., in an answer to *What did he pick up?*. (36) is an ordering corresponding to one of three possible situations: (i) 'pen' is CB and 'up' is NB, e.g. in an answer to *What did he do with the pen?*, (ii) both 'pen' and 'up' being NB, e.g. in an answer to *What did he do?*, (iii) both 'pen' and 'up' being CB, e.g. in an answer to *Who picked the pen up?*, and with a marked intonation stressing 'he'.

In FGD, the CB/NB distinction is also held responsible for the so-called **dative shift** in English. Whereas (37) is considered to be neutral, for (38) it is said that 'Christmas cards' are understood as CB—presuming the neutral intonation patterns with the intonation center on the last sentence element (Sgall *et al.* 1986: p.201):

(37) A young linguist sent a few colleagues Christmas cards.

(38) A young linguist sent Christmas cards to a few colleagues.

The view that the ordering of the Participants corresponding to the Actor, Patient and Addressee is constrained by information structure in the same way as the ordering of Circumstances, is natural for Czech,

[10] For a "phrasal verb", it is quite likely that the particle's CB/NB assignment follows the CB/NB assignment of the verb. In order to pursue this line further, we would first have to be more specific regarding the treatment of CB/NB for verbs and complex verb forms, and its influence on CD and consequently on WO. Such discussion is, however, beyond the scope of the present paper.

where the Participants are realized by nominal groups in different morphological cases and their order is free.

Adopting such an account also for English would mean that information structure would determine the grammar's choice between (i) realizing both the Patient and the Addressee by a nominal group in the oblique case, or (ii) realizing the Patient as a direct complement and the Addressee as a prepositional phrase. Such an approach, however, may need to be refined. For example, Steele (1978) argued that there is a difference in meaning between (39) and 40):

(39) Mary threw John the ball. (40) Mary threw the ball to John.

Steele argues here that in (39) John has to be actively participating in the act, whereas in (40) he need not. Hence, she considers (39′) semantically awkward, while (40′) would be fine.

(39′) # Mary threw John the ball but he wasn't looking.

(40′) Mary threw the ball to John, but he wasn't looking.

We therefore leave it open at this moment whether dative shift should be attributed solely to information structure or also to a meaning difference (see also (Oehrle 1975), (Kunze 1991)).

Also Czech WO is subject to grammatical restrictions, although to a lesser degree than English WO. Besides the ordering of clitics (Section 1.2, p. 5) there is also a preference for **verb second placement**. This is what explains the difference in WO between (31) and (41):

(41) Ikonu přeneseme myší do
$\text{icon}_{acc}{}^{Patient-CB}$ drag_{1pl} $\text{mouse}_{ins}{}^{Means-CB}$ into
editoru z plochy.
$\text{editor}_{gen}{}^{Destin-CB}$ from $\text{desktop}_{gen}{}^{Source-NB}$
The icon we drag with the mouse to the editor from the desktop.

Both in English and Czech, there are also orders we cannot explain solely on the basis of the CB/NB distinction along with SO and grammatical constraints. On the one hand, it has been claimed that the ordering of CB elements follows CD rather than SO, and that CD is determined by contextual factors (Sgall *et al.* 1986). On the other hand, cases where an NB element appears at the beginning of a clause are far from rare. While we currently do not have more to add to the former issue, we will discuss the latter issue in the next section. For illustration, consider (42) used as the first step in a sequence of instructions for editing an existing file. The context can thus be illustrated by the question *What is to be done with the file?*.

(42) Myší přeneseme ikonu z
Mouse$_{ins}{}^{Means-NB}$ drag$_{1pl}$ icon$_{acc}{}^{Patient-CB}$ from
plochy do editoru.
surface$_{gen}{}^{Source-NB}$ into editor$_{gen}{}^{Destin-NB}$

Using the mouse, we drag the icon from the desktop to the editor.

For the sake of argument, we can consider 'icon' as CB (making such decision in the course of text planning would be justified by the fact that the icon of a file can be assumed to be known to the reader of the instruction text). The placement of this CB element after rather than before the verb in the Czech version of (42) can be explained by verb secondness. However, how do we explain the placement of 'mouse' at the beginning of the sentence?

In the Prague School tradition, it has always been acknowledged that WO can diverge from the prototypical ordering proceeding left-to-right from CB (Given) to NB (New) elements. By inverting the order, and presenting the NB (New) elements first, one obtains what has been called the **subjective order** (Firbas 1992). Example (30′) illustrates the subjective-order variant of (30), which could be used to answer the same question, i.e., *What do we drag from the desktop to the editor using the mouse?*,

(30′) Ikonu přeneseme myší z
icon$_{acc}{}^{Patient-NB}$ drag$_{1pl}$ mouse$_{ins}{}^{Means-CB}$ from
plochy do editoru.
desktop$_{gen}{}^{Source-CB}$ into editor$_{gen}{}^{Destin-CB}$

The icon we drag with the mouse from the desktop to the editor.

The speaker has the choice of answering the question using the unmarked prototypical order or the marked subjective order. Note that the subjective order needs to be accompanied with a marked intonation pattern stressing 'icon' (in both Czech and English). It is probably due to its close coupling to intonation that the subjective order is more common in spoken language than in writing. We did not encounter any occurrences of the subjective order in our corpus of written instructions. Therefore, we do not consider it further in the present account.

However, even subjective order does not enable us to explain the WO in (42). First of all, in line with all other examples, (42) is intended to be pronounced with unmarked intonation, i.e. with the intonation center on 'editor'. Moreover, if subjective order were used, then all NB elements should occur before the CB ones, which is not the case in (42). So, we conclude that we cannot motivate the WO in (42) on the basis of information structure (alone).

1.4.2 Theme and Textual Organization

Central to Halliday's approach is the notion that many languages *grammaticize* a very particular textual function: that of signposting the intended development or "scaffolding" that a writer employs for structuring an extended text. In the grammar of English this textual function is realized in the first position in the clause and Halliday calls this textual function 'Theme': "The Theme is the element which serves as the point of departure of the message; it is that with which the message is concerned. [...] and the [thematic] structure is expressed by the order - whatever is chosen as the Theme is put first" (Halliday 1985: p.37).

Halliday's use of the term 'Theme', which we henceforth abbreviate as **Theme**$_{SFG}$, can only be understood in relation to the thematic structure. The specialized textual role of Theme$_{SFG}$ was originally identified empirically in the work of Fries (1981) and has since been applied in many analyses of extended texts (e.g., (Halliday 1993; Matthiessen 1995; Ventola 1995)). Although the claim of this special role for the first position in the clause has been questioned or rejected in various works (Lyons 1977; Huddleston 1988; Bolinger 1989; Steedman 1996, to appear), these critiques miss the point that it is the larger **textual organization** to which Theme$_{SFG}$ contributes, not the local sentence-by-sentence organization found in dialogue or as highlighted by question tests. For instance, Steedman's notion of 'Theme', which we henceforth abbreviate as **Theme**$_{is}$ diverges from Theme$_{SFG}$ as follows:[11] (i) Theme$_{is}$ does not need to be sentence-initial, it can be ordered before or after Rheme$_{is}$; (ii) Theme$_{is}$ can contain multiple experiential elements (for example, multiple circumstantial modifiers, but also multiple participants, as well as the main verb), and (iii) disjoint parts of a sentence might belong to Theme$_{is}$ (Steedman to appear: p.7). It is an unfortunate terminological confusion that most non-Hallidayan uses of 'Theme' refer instead to properties that Halliday attributes to information structure.

Theme$_{SFG}$ is thus concerned primarily with textual organization. For example, it is a matter of such global text organization strategy that a biography can be organized along the dates when or the places where important events in one's life happened. These dates or places are not "given" information individually, but they are nevertheless broadly "predictable" from the knowledge of the text type. Such strategies then guide the choice of Theme.

Similar textual organization strategies are also at play in the written instructions for software products. For example, consider (44) which

[11] Similar comments apply to the Theme$_{is}$ counterparts in other approaches to information structure, e.g., Topic (Sgall *et al.* 1980, 1986) or Ground (Vallduví 1992).

demonstrates a kind of thematization appearing repeatedly in the English, Czech, Bulgarian and Russian versions of the AutoCAD user guide. This particular occurrence was preceded only by the respective language counterpart of (43).

(43) First open the Multiline styles dialog box using one of the following methods.

(44) **Cz** Z menu `Data` vyberte `Multiline Style`.
Ru В меню `Data` выберите `Multiline Style`
Bu От менюто `Data` изберете `Multiline Style`
En From the `Data` menu choose the `Multiline Style`.

The preceding context does not refer to the `Data` menu or make it salient in any way. Working only with the notion of information structure discerning CB (Given) and NB (New) elements, one is thus unable to explain the WO. The notion of Theme$_{SFG}$ as a reflex of a global text organization strategy makes such explanation possible for (44) as well as for (42) in the preceding section, p. 16.

In the software instructions we analyzed, we encountered regular thematization of (i) the location where actions are performed (44,45), (ii) the particular action that the user is instructed to perform (46), or (iii) the goal that the user wants to achieve (47).[12]

(45) Z menu `File` vyberte `Save` pro uložení souboru.
from menu$_{gen}$ `File` choose$_{i2pl}$ `Save` for saving$_{acc}$ file$_{gen}$
"From the `File` menu, choose `Save` to save the file."

(46) Vyberte `Save` z menu `File` pro uložení souboru.
choose$_{i2pl}$ `Save` from menu$_{gen}$ `File` for saving$_{acc}$ file$_{gen}$
"Choose `Save` from the `File` menu to save the file."

(47) Pro uložení souboru vyberte `Save` z menu `File`.
for saving$_{acc}$ file$_{gen}$ choose$_{i2pl}$ `Save` from menu$_{gen}$ `File`
"To save the file, choose `Save` from the `File` menu."

If we cannot avail ourselves of Theme$_{SFG}$, then rather more of the WO decisions in an extended text appear unmotivated than need be; this has been shown with respect to examples drawn from extended narrative texts in, for example, (Martin 1995). For a text generation system this would be particularly unfortunate; the rather straightforward selection of initial constituents in many text types would then not be available. It is for this role that we maintain the use of Theme$_{SFG}$ for text generation.

Conversely, we also certainly do not want to deprive ourselves of the constraining power for the *rest* of the clause offered by information structure. The position we advance, therefore, is to distinguish between information structure and the textually determined thematic structure, and

[12]The same variation of thematic structure is available also in Bulgarian and Russian.

employ *both* of them as orthogonally applying sources of WO constraint. By this, we follow the original suggestion made by Halliday (1967).

1.5 Word Order Control in Generation

Our approach to handling WO fits within the standard "end-to-end" architecture of a multilingual NLG system. Such a system takes as an input user-defined (language-independent) content specifications. These are first processed by a text and sentence planner (strategic generation), yielding a list of formulae in a **sentence planning language** (SPL) formulae (Kasper 1989). These SPL formulae are then processed by the grammar(s) generating the individual sentences in each target language (tactical generation).

It is in this setting, in which individual sentences are specified and generated that ultimately make up a text, that the issue of generating contextually appropriate sentences arises. For only in that way can a smooth flow of text be assured.

1.5.1 Flexible Word Order Algorithm

As we have discussed, constraints from various sources need to be combined in order to determine grammatically well-formed and contextually appropriate WO. We propose to ensure this as follows:

- The grammar imposes structural constraints, such as the ordering of connectives, placement of sentential clitics and possibly verb secondness.
- The text planner can choose a particular textual organization and determine the element which should become the Theme. If no particular element is chosen as the Theme, the grammar chooses the least communicatively dynamic element as the Theme. The Theme is then placed at the beginning of the clause, although not necessarily at the very first position, as this might be occupied, e.g., by a connective. The placement of the Theme can also be resolved by the grammar.
- The text and sentence planner determine information structure, in particular, contextual boundness or non-boundness of each item. During text planning, a discourse model is maintained which records the use of pieces of content. Whenever the sentence planner encounters a piece of content that the discourse model notes as previously used, it marks the corresponding item in the SPL formula as CB.
- Contextual boundness is then used to constrain WO (at the clause level) as follows:

- The CB elements (if there are any) precede the NB elements in the prototypical cases, that is, leaving the subjective order aside.
- The mutual ordering of multiple CB items in a clause corresponds to communicative dynamism, and the mutual ordering of multiple NB items in a clause follows the systemic ordering (with the exceptions required by grammatically constrained ordering as described below). The default for communicative dynamism is systemic ordering.
- The main verb of a clause is ordered at the boundary between the CB elements and the NB elements, unless the grammar specifies otherwise (verb secondness).

We combine the following two phases in which information structure (CB/NB) is taken into account during tactical generation:

- information structure can determine particular realization choices made in the grammar; for example, when inserting and placing the particle of a phrasal verb, when inserting and ordering the Source and Destination for a motion process (cf. Section 1.4.1, p. 14), and possibly also when ordering elements at lower structural levels, e.g., within nominal groups, etc.
- information structure can determine the ordering of elements whose placement has not been sufficiently constrained by the grammar

This also enables each language within multilingual resources to establish its own balance between the two phases. To show our approach in a nutshell, we present an abstract WO algorithm in Figure 1.

The ordering constraints posed by the grammar have the highest priority. Note that this includes the ordering of the textually determined Theme. Then, any elements which are not ordered by the grammar, are subject to the ordering according to information structure, i.e. systemic ordering in combination with the CB/NB distinction. The ordering of the NB elements (i) restricted by the syntactic structure or (ii) follow SO. The ordering of the CB elements can be (i) specified on the basis of the context, (ii) restricted by the syntactic structure, (iii) follow SO. The verb is placed between the last CB and the first NB element, unless its placement has been determined by the grammar, e.g. if is itself the Theme or due to verb secondness.

The ordering algorithm as such is language independent, and could be usefully applied in the generation of any language. What differs across languages is first of all the extent to which the grammar of a particular language constrains ordering, i.e. which elements are subject to ordering requirements posed by the syntactic structure and which elements

FIGURE 1 Abstract ordering algorithm

```
Given:
    a set GC of ordering constraints imposed by the grammar
    a list L1 of constituents that are to be ordered,
    a list D giving ordering of CB constituents (default is SO)

Create two lists LC and LN of default orders:

    Create empty lists LC (for CB items) and LN (for NB items)
    Repeat for each element E in L1
            if E is CB,
                then add E into LC,
                else add E into LN.
    Order all elements in LC according to D
    Order all elements in LN according to SO
    if the Verb is yet unordered then
        Order the Verb at the beginning of LN

Order the elements of L1
    if GC is not empty then
         use the contraints in GC, and
         if the contraints in GC are insufficient,
            apply first the default orders in LC and then those in LN
```

can be ordered according to information structure. Secondly, individual languages may differ in the SO (cf. Section 1.4.1, p. 11).

1.5.2 Algorithm Application

The algorithm described above has been implemented and used in the process of tactical generation of Czech and English. The Czech grammar resources used at this stage in generation have been built up along with the Bulgarian and Russian grammar resources as described by Kruijff *et al.* (2000). The NIGEL grammar for English was re-used for this purpose.

The original NIGEL grammar itself already combines the specification of ordering constraints in the grammar with the application of defaults. In any ordering is underspecified in the grammar, the defaults are applied. The defaults are "static", i.e. specified once and for all. The algorithm we have described can be used to replace these "static" defaults with a "dynamic" construction of default ordering constraints. Two separate sets of "dynamic" defaults are computed on the basis of the SO for the CB and the NB elements in each sentence/clause.

We use the SO specified in (26) for Czech and the SO specified in (27) for English (cf. Section 1.4.1, p.11). For each element in the input SPL we specify whether it is CB (`:contextual-boundness yes`) or NB (`:contextual-boundness no`); in addition, we can specify the textual

Theme in the SPL (`theme <id>`). The SPL in Figure 2 illustrates this (the corresponding generated output in shown in (53)).

The CB/NB assignments can be varied to obtain different WO variants (cf. Section 1.4.1). The examples below list some of the CB/NB assignment combinations and the outputs generated using the Czech and English grammars.

(48)

user	choose	`Open`	menu	mouse	open file
Actor-NB		Patient-NB	SpaceLoc.-NB	Means-NB	Purpose-NB

Uživatel	vybere	pro otevření	souboru	v menu	myší	`Open`.
user$_{nom}$	choose$_{3sg}$	for opening$_{gen}$	file$_{gen}$	in menu$_{loc}$	mouse$_{ins}$	`Open`

The user chooses `Open` in the menu with the mouse to open a file.

(49)

user	choose	`Open`	menu	mouse	open file
Actor-CB		Patient-NB	SpaceLoc.-CB	Means-NB	Purpose-NB

Uživatel	v menu	vybere	pro otevření	souboru	myší	`Open`.
user$_{nom}$	in menu$_{loc}$	choose$_{3sg}$	for opening$_{gen}$	file$_{gen}$	mouse$_{ins}$	`Save`

The user chooses `Open` in the menu with the mouse to open a file.

(50)

user	choose	`Open`	menu	mouse	open file
Actor-CB		Patient-NB	SpaceLoc.-CB	Means-CB	Purpose-CB

Pro otevření	souboru	uživatel	v menu	myší	vybere	`Open`.
for opening$_{gen}$	file$_{gen}$	user$_{nom}$	in menu$_{loc}$	mouse$_{ins}$	choose$_{3sg}$	`Open`

To open a file the user chooses `Open` in the menu with the mouse .

As we said above, we preserve the notion of textual Theme. An SPL can contain a specification of a theme, and the corresponding element is then ordered at the front of the sentence, as determined by the grammar. The WO of the rest of the sentence is determined as before. The following examples of generated output illustrate this.

(51)

user	choose	`Open`	menu	mouse	open file
Actor-NB		Patient-NB	SpaceLoc.-NB	Means-NB	Purpose-theme

Pro otevření	souboru	uživatel	vybere	v menu	myší	`Open`.
for opening$_{gen}$	file$_{gen}$	user$_{nom}$	choose$_{3sg}$	in menu$_{loc}$	mouse$_{ins}$	`Open`

To open a file, the user chooses `Open` in the menu with the mouse.

(52)

user	choose	`Open`	menu	mouse	open file
Actor-CB		Patient-NB	SpaceLoc.-CB	Means-CB	Purpose-CB

Pro otevření	souboru	uživatel	v menu	myší	vybere	`Open`.
for opening$_{gen}$	file$_{gen}$	user$_{nom}$	in menu$_{loc}$	mouse$_{ins}$	choose$_{3sg}$	`Open`

To open a file, the user chooses `Open` in the menu with the mouse .

FIGURE 2 The input SPL for both the English and the Czech version of (53)

```
(R / RST-purpose
    :speechact assertion
    :DOMAIN (ch / DM::choose
        :actor (a1 / DM::user
            :identifiability-q identifiable
            :contextual-boundness yes)
        :actee (a2 / object :name gui-open
            :identifiability-q identifiable
            :contextual-boundness no)
        :instrumental (mea / DM::mouse
            :identifiability-q identifiable
            :contextual-boundness no)
        :spatial-locating (loc / DM::menu
            :identifiability-q identifiable
            :contextual-boundness yes
            :class-ascription (label / object :name gui-file))
      :RANGE (open / DM::open
          :contextual-boundness no
          :actee (f / DM::file :contextual-boundness no)))
      :theme open)
```

(53) user choose `Open` menu mouse open file
Actor-CB Patient-NB SpaceLoc.-CB Means-NB Purpose-theme

Pro otevření souboru uživatel v menu vybere myší **Open.**
user$_{nom}$ in menu$_{loc}$ choose$_{3sg}$ for opening$_{gen}$ file$_{gen}$ mouse$_{ins}$ **Save**

To open a file, the user chooses `Open` in the menu with the mouse.

In both the English and Czech grammars we are using, there are already several decision points whose control naturally calls for access to information structure. These include the particle placement and dative shift in English (mentioned in Section 1.4.1, p. 14), as well as components of the active/passive and pronominalization decisions. Given the functional orientation of our grammars, it would be a natural move to take the CB/NB distinction into account when making these choices.

Our last note concerns handling clitics in Czech. There are essentially two options for the implementation: (i) to introduce a specially handled category for clitics in the ordering algorithm, (ii) to order clitics explicitly in the grammar. In the current version of the Czech grammar we opt for the latter. Using < for linear precedence, the ordering we implemented for simple clauses can be schematized as follows:

`Theme` < `Clitics` < `Rest-CB` < `Verb` < `Rest-NB`

If no Theme has been determined by the text planner, then the Czech grammar takes the CB element with the lowest SO degree as the Theme;

if there is no CB element, then the verb becomes the Theme. The examples of generated Czech output below illustrate this.

(54) choose `Open` menu mouse open file
Patient-NB SpaceLoc.-NB Means-NB Purpose-NB

Vybere se pro otevření souboru v menu myší `Open`.
choose_{3sg} *refl* for opening_{gen} file_{gen} in menu_{loc} mouse_{ins} `Open`

"One chooses `Open` in the menu with the mouse to open a file."

(55) choose `Open` menu mouse open file
Patient-NB SpaceLoc.-CB Means-NB Purpose-NB

V menu se vybere pro otevření souboru myší `Open`.
in menu_{loc} *refl* choose_{3sg} for opening_{gen} file_{gen} mouse_{ins} `Save`

"One chooses `Open` in the menu with the mouse to open a file."

(56) choose `Open` menu mouse open file
Patient-NB SpaceLoc.-CB Means-CB Purpose-theme

Pro otevření souboru se v menu myší vybere `Open`.
for opening_{gen} file_{gen} *refl* in menu_{loc} mouse_{ins} choose_{3sg} `Open`

"To open a file one chooses `Open` in the menu with the mouse."

Thus, our approach combines WO constraints imposed by syntactic structure, thematic structure and information structure. Information structure is determined by the CB/NB assignments in the input SPL. The language-specific SO is employed as a source of ordering defaults.

1.6 Concluding Remarks and Further Research

We have explained the main factors determining word order and presented a flexible approach to handling word order in the course of automatic "end-to-end" natural language generation. Our approach on the one hand enables us to cover all the grammatically well-formed ordering possibilities available in a given language. On the other hand, it makes it possible to control the choice of contextually appropriate variants.

We employ the SFG notion of a textually determined Theme, because it appears necessary in any language to account for text structuring concerns across sentences within connected spans of texts. For instance, the decision what to chose as a "point of departure" can be motivated by a particular style chosen for the text, in which case one should not need to look for a motivation for a particular ordering based on information structure.

For ordering of non-thematic constituents within a clause, when their ordering is not determined by the syntactic structure, we use notions adopted from FGD, namely the distinction between contextual boundness and non-boundness in combination with the so-called systemic ordering.

This integration of the SFG and FGD insights concerning ordering enables us to formulate a sufficiently flexible as well as sufficiently constrained approach to handing word order, suitable for languages with varying degree of word order freedom.

We believe that the discussion we have provided raises some general issues concerning system modularity and architecture. The focus of the workshop where an earlier version of this paper was presented was on the generation of nominal expressions. In this respect, there is the question of what should be considered as the task of nominal expression generation. Although a first approximation to nominal expression generation might posit a 'nominal expression generation module' whose responsibility it is to map chunks of semantic/conceptual representation into appropriate nominal expressions, our discussion in this paper demonstrates that the boundaries in semantic responsibility within an NLG architecture are far from clear. Given such a module, the ordering of clause elements might seem to lie outside rather than inside the module as such. However, factors that play a role in determining word order in Slavonic languages are often precisely the same factors that play a role in determining certain syntactic, morphological or phonological features of/within nominal expressions in other languages. It would seem that some subset of the decisions that have been proposed for generating nominal expressions may instead be better considered as decisions concerning *clause* generation. This may also occur when we consider other components of the generation process. Therefore, architectural decisions made on the basis of surface syntactico-morphological behaviour of any individual language may well turn out not to be appropriately general when other languages come to be considered.

Acknowledgements The paper builds on work we have carried out in relation to the AGILE project (Automatic Generation of Instructions in Languages of Eastern Europe) supported by the European Commission grant No. PL961104. The overall aim of AGILE was to develop a multilingual system for generating continuous instructional texts in Bulgarian, Czech and Russian (Adonova *et al.* 2000). We would like to thank Elke Teich, Tony Hartley, Eva Hajičová and two anonymous reviewers of earlier versions of this paper for their comments; Serge Sharoff and Lena Sokolova contributed examples and judgments for Russian, Danail Dochev and Kamenka Staykova for Bulgarian.

References

Adonova et al. 1999. E. Adonova, J. Bateman, N. Gromova, A. Hartley, G. J. M. Kruijff, I. Kruijff-Korbayová, S. Sharoff, H. Skoumalová, L. Sokolova, K. Staykova, and E. Teich: Formal specification of extended grammar models. AGILE project deliverable. Technical Report. ITRI, University of Brighton, UK.

Bolinger 1989. D. Bolinger. *Intonation and its Uses.* Stanford University Press.

Daneš 1985. F. Daneš. *Věta a text [A sentence and a text]*, volume 21 of *Studie a práce lingvistické.* Academia, Prague, Czech Republic.

Firbas 1992. J. Firbas. *Functional Sentence Perspective in Written and Spoken Communication.* Studies in English Language. Cambridge University Press.

Fries 1981. P. H. Fries. On the status of theme in English: Arguments from discourse. *Forum Linguisticum*, 6(1):1 – 38.

Gries 1999. S. T. Gries. Particle movement: a cognitive and functional approach. *Cognitive Linguistics*, 10(2):105–146.

Hajičová 1993. E. Hajičová. *Issues of sentence structure and discourse patterns*, volume 2 of *Theoretical and computational linguistics.* Charles University, Prague, Czech Republic.

Hajičová and Kruijff-Korbayová 1999. E. Hajičová, and I. Kruijff-Korbayová. On the notion of topic. In *Prague Linguistic Circle Papers.* John Benjamins Publishers, Amsterdam, the Netherlands.

Hajičová and Sgall 1987. E. Hajičová, and P. Sgall. The ordering principle. *Journal of Pragmatics*, 11(4):435–454.

Hajičová et al. 1995. E. Hajičová, P. Sgall, and H. Skoumalová. An automatic procedure for topic-focus identification. *Computational linguistics*, 21(1):81–94, 1995.

Halliday 1967/8. M. A. K. Halliday. Notes on transitivity and theme in English. *Journal of Linguistics*, 3(1):37–81, 3(2):199–244, 4(2):179–215.

Halliday 1993. M. A. K. Halliday. The construction of knowledge and value in the grammar of scientific discourse: Charles Darwin's *The Origin of the Species.* In M. A. K. Halliday and J. R. Martin, editors, *Writing Science: literacy and discoursive power*, pages 86–105. The Falmer Press, London.

Halliday 1970 . M. A. K. Halliday. *A Course in Spoken English: Intonation.* Oxford Uniersity Press, Oxford.

Halliday 1985. M. A. K. Halliday. *Introduction to Functional Grammar.* Edward Arnold, London, U.K.

Hudleston 1988. R. D. Huddleston. Constituency, multi-functionality and grammaticalization in Halliday's Functional Grammar. *Journal of Lin-*

guistics, 24.

Kasper 1989. R. Kasper. A flexible interface for linking applications to PENMAN's sentence generator. In *Proceedings of the DARPA Workshop on Speech and Natural Language.*

Krifka 1992. M. Krifka. A compositional semantics for multiple focus constructions. In J. Jacobs, editor, *Informationsstruktur und Grammatik*, Linguistische Berichte, Sonderheft 4.

Kruijff et al. 2000. G. J. M. Kruijff, E. Teich, J. Bateman, I. Kruijff-Korbayová, H. Skoumalová, S. Sharoff, L. Sokolova, A. Hartley, K. Staykova, and J. Hana. A multilingual system for text generation in three slavic languages. In *Proceedings of the 18th Conference on Computational Linguistics (COLING'2000), July 31 - August 4 2000*, pages 474–480. Universität des Saarlandes, Saarbrücken, Germany.

Kruijff forthcoming. G. J. M. Kruijff. *A Categorial-Modal Logical Architecture of Informativity: Dependency Grammar Logic & Information Structure.* PhD thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

Kuno 1972. S. Kuno. Functional sentence perspective. *Linguistic Inquiry*, 3:269–320.

Kunze 1991. J. Kunze. *Kasusrelationen und semantische Emphase.* Studia Grammatica XXXII. Akademie Verlag, Berlin.

Lyons 1977. J. Lyons. *Semantics*, volume II. Cambridge University Press.

Martin 1995. J. R. Martin. More than what the message is about: English theme. In Mohsen Ghadessy, editor, *Thematic development in English texts*, pages 223–258. Pinter, London.

Mathesius 1939. V. Mathesius. On information bearing structure of the sentence. In S. Kuno, editor, *Harvard studies in syntax and semantics.* Harvard University Press. 1975.

Mathiessen 1995. C. M. I. M. Matthiessen. Theme as an enabling resource in ideational 'knowledge' construction. In M. Ghadessy, editor, *Thematic development in English texts*, pages 20–54. Pinter Publishers, London.

Oehrle 1975. R. T. Oehrle. *The Grammatical Status of the English Dative Alternation.* PhD thesis, M.I.T.

Oliva and Avgustinova 1995. K. Oliva, and T. Avgustinova. Wackernagel position and related phenomena in Czech. In *Wiener Slavistiches Jahrbuch*, pages 21–42. Verlag der Österreichischen Akademie der Wissenschaften, Wien.

Panevová 1974. J. Panevová. On verbal frames in functional generative description I. *Prague Bulletin of Mathematical Linguistics*, 22:3–40.

Petkevič 1987. V. Petkevič. A new dependency based specification of underlying representations of sentences. *Theoretical linguistics*, 14:143–172. Walter de Gruyter, Berlin, Germany.

Petkevič 1995. V. Petkevič. A new formal specification of underlying structures. *Theoretical linguistics*, 21(1):7–61. Walter de Gruyter, Berlin, Germany.

Prevost 1995. S. Prevost. *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*. Ph.d. dissertation, IRCS TR 96-01, University of Pennsylvania, Philadelphia.

Prevost and Steedman 1994. S. Prevost, and M. Steedman. Specifying intonation from context for speech synthesis. *Speech Communication*, 15:139–153.

Sgall et al. 1973. P. Sgall, E. Hajičová, and E. Benešová. *Topic, focus and generative semantics*. Kronberg, Taunus.

Sgall et al 1980. P. Sgall, E. Hajičová, and E. Buráňová. *Aktuální členění věty v češtině (Topic-focus articulation of the Czech sentence)*. Academia, Prague, Czech Republic.

Sgall et al. 1986. P. Sgall, E. Hajičová, and J. Panevová. *The meaning of the sentence in its semantic and pragmatic aspects*. Reidel, Dordrecht, The Netherlands.

Sgall et al. 1995. P. Sgall, O. Pfeiffer, W. U. Dressler, and M. Půček. Experimental research on systemic ordering. *Theoretical Linguistics*, 21:197–239.

Steedman 1994. M. Steedman. The grammar of intonation and focus. In Paul Dekker and Martin Stokhof, editors, *Proceedings of the 9th Amsterdam Colloquium*. Institute of logic, language and computation, Department of Philosophy, Universiteit van Amsterdam.

Steedman 1996. M. Steedman. *Surface Structure and Interpretation*. M.I.T. Press.

Steedman 2000. M. Steedman. *The Syntactic Process*. M.I.T. Press.

Steedman to appear. M. Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*.

Steele 1978. S. Steele. Word order variation: A typological study. In Joseph H. Greenberg, editor, *Universals of Language. Volume 4: Syntax*, pages 585–624. Stanford University Press.

Vallduví 1992. E. Vallduví. *The Informational Component*. Garland, New York.

Ventola 1995. E. Ventola. Thematic development and translation. In M. Ghadessy, editor, *Thematic development in English texts*, pages 85–104. Pinter Publishers, London.