Contextually Appropriate Ordering of Nominal Expressions*

Ivana Kruijff-Korbayová   Geert-Jan Kruijff

Contents

*This paper is based on research carried out within the AGILE project. AGILE (Automatic Generation of Instructions in Languages of Eastern Europe) is an international project supported by the European Commission within the COPERNICUS programme, grant No. PL961104. The overall aim of the AGILE project is to develop a multilingual system for generating continuous instructional texts in Bulgarian, Czech and Russian [12]. The project is concerned mainly with (i) the development and adaptation of linguistic resources for the chosen languages, and (ii) the investigation and specification of text structuring strategies employed in those languages for the given type of texts. The linguistic resources are developed in the KPML environment [3], on the basis of the Penman system [2]. We would like to thank our colleagues from the Charles University, University of Brighton, Bulgarian Academy of Sciences, Russian Research Institute of Artificial Intelligence, and Elke Teich and John Bateman in particular, for their cooperation. We are also grateful to the reviewers of the abstract who provided useful comments which we tried to use to improve the presentation of our ideas.

1 Introduction

Slavonic languages exhibit a relatively high degree of word order freedom. This characteristic is based on their comparison with languages like English or French, where clause constituents cannot be "moved around" with the same relative freedom without simultaneous changes in the syntactic structure. However, various word order variants of a sentence, even though they are grammatically well-formed, do not necessarily have the same meaning and are generally not interchangeable in a given context. This means that in the process of automatic generation of continuous texts from an underlying representation of the content, we have to ensure that a semantically and contextually appropriate word order is chosen. Even in English, a lack of a proper account of the ordering phenomena yields a less fluent generated output.

In this paper, we address the relevance of word order in the course of automatic natural language generation. The main questions we address are as follows:

- What differences in meaning are conveyed by different word order in a written text in Czech, Russian and Bulgarian?
- What are the factors that determine word order in these languages?
- How can word order in the course of multilingual generation be controlled?

We concentrate in particular on the ordering of nominal expressions, i.e. nominal groups and prepositional phrases. We consider both their mutual ordering within clauses, and their ordering with respect to the main verb. We elucidate the main factors determining word order in Slavonic languages, and present an approach to handling word order in a system of automatic text generation we are developing in the context of the AGILE project.

In order to account for the word ordering phenomena within the AGILE project, we build on the insights of existing linguistic theories. Currently we restrict ourselves to combining the following two approaches:

- Halliday's thematic structure [9] as developed in the Systemic Functional Grammar framework (SFG) is chosen because SFG is the framework adopted in the Penman system [2] on which our AGILE grammars, developed in the KPML environment [3], are based;
- the topic-focus articulation approach developed within the framework of Functional Generative Description (FGD, [15]) serves us to elaborate the SFG approach towards a more flexible treatment required for languages with a higher degree of free word order than English, especially because Halliday's approach is not sufficiently specific with respect to the ordering of non-thematic constituents.[1]

The paper consists of two main parts. In the first part, we discuss the relevance of word order phenomena. We begin by an illustration of the differences

[1] For a detailed comparison of the SFG and FGD approaches see [1].

in semantic and contextual appropriateness of simple sentences with varying word order in the three languages (Section 2.1). Then we briefly overview the essential theoretical notions of SFG and FGD which are relevant for handling word order (Section 2.2).

In the second part, we describe our approach, which enables us to determine a suitable word order when generating a sentence conveying a particular content in a given language in a particular context. We propose a way in which to combine the insights of SFG and FGD concerning word order, and formulate the corresponding ordering algorithm (Section 3.1). Our word ordering algorithm operates on the elements in a clause. In order to show its role within the entire process of text generation which takes as an input a knowledge representation of the content to be conveyed by the text, we describe our text planning strategy and the interface between the text planner and the sentence generator (Section 3.2).

We conclude the paper with a discussion of the overall coverage of the proposed approach with respect to the contextually appropriate realization of information structure by word order (Section 4).

2 The Relevance of Word Order Variation

In natural language communication, each participant can only process the elements one by one, in a linear order. However, the content that is being communicated is by nature multi-dimensional. Linguistic structures of all degrees of complexity therefore need to be projected into one dimension, and the individual elements ordered sequentially according to certain rules. The rules and schemes of linear ordering can be studied at every stratum of a given language, including the lowest ones, i.e. graphemes, phonemes, morphemes etc. The rules and schemes pertaining to the linear ordering of elements constituting a clause, i.e. phrases and groups, are usually referred to as *word order* (even though the elements are not only individual words). The linear ordering of clauses within a complex clause (sentence) can be referred to as *clause order*.

Various factors can be discerned in the language system in general that play an important role in expressing a given content in a linear form. The inventory of these factors contains at least the following: *information structure, grammatical structure, intonation, rhythm* and *style*. These factors are very general, and can therefore be considered language universals, at least within the family of indo-european languages. However, the individual factors may have different importance for the linear ordering, i.e. word order and clause order, in a given language.

For instance, English is an example of a language where word order is strongly constrained by grammatical structure. In such a language with a rather fixed word order, differences in information structure are reflected by varying the intonation pattern or by the choice of a particular type of expression. The latter concerns, e.g., the use of a definite vs. indefinite nominal group to refer anaphorically vs. to introduce new entities, respectively. In languages such as Czech, the same effects are achieved by varying word order in accordance to information structure.

Below, we first discuss some examples of word order variation, and then briefly present the essential theoretical notions employed in our account of word order in the AGILE project.

2.1 Discussion of Examples

For sentences with a moderate number of syntactic groups there are usually quite a few grammatically well formed word order variants in Slavonic languages. However, one should not interpret the high degree of word order freedom as its arbitrariness.

For illustration, consider the following examples of word order variation. The sentences contain one main verb and two or more nominal groups or prepositional phrases appearing either before or after the verb in different orders.[2] They are modifications of sentences from the instruction texts generated in the AGILE project. The target texts are adapted from a CAD/CAM system user guide.

(1)	**Cz**	Otevřete	soubor	příkazem	*Open.*
	Ru	Otkrojte	fajl	kommandoj	*Open.*
	Bu	Otvorete	fajla	s komandata	*Open.*
	gl	Open-imp	file-acc	command-instr	*Open*

En Open a file by the *Open* command.

(2)	**Cz**	Soubor	otevřete	příkazem	*Open.*
	Ru	Fajl	otkrojte	kommandoj	*Open.*
	Bu	Fajla	otvorete	s komandata	*Open.*
	gl	File-acc	open-imp	command-instr	*Open.*

En Open the file by the *Open* command.

(3)	**Cz**	Příkazem	*Open*	otevřete	soubor.
	Ru	Kommandoj	*Open*	otkrojte	fajl.
	Bu	S komandata	*Open*	otvorete	fajla.
	gl	Command-instr	*Open*	open-imp	file-acc

En By the *Open* command, open a file.

(4)	**Cz**		Příkazem	*Open*	soubor	otevřete.
	Ru	⋆	Kommandoj	*Open*	fajl	otkrojte.
	Bu	⋆	S komandata	*Open*	fajla	otvorete.
	gl		Command-instr	*Open*	file-acc	open-imp

En By the *Open* command, open the file.

[2] We only present one common gloss, though there are some differences between the languages. Czech and Russian have richer morphology, and distinguish seven and six cases, respectively; Bulgarian only differentiates between four cases. In the discussed example, instrument is expressed by instrumental case in Czech and Russian, but it is expressed using the preposition 's' in Bulgarian. Another difference concerns the use of articles: whereas Czech and Russian do not have a definite or indefinite article, Bulgarian does have a definite article, which appears as the suffix '-ta' in 'komandata' in the discussed example.

The following judgments hold in all three languages. The ordering in the sentence in (1) is neutral. It can be used "out of the blue", or in a context which can be approximated by the question *What should we do?* There are no implicatures concerning the presence of a file, or the identity of the file to be opened.

The sentence in (2) means opening of a specific file. It is appropriate when some file is salient, for instance when the user is working with a file. That is why we put the definite article into the English translation. The action of opening can, but does not have to, be salient, too. The contexts in which (2) can be appropriately used can be characterized by the questions *What should we do with the file?* or *How should we open the file?*.

(3) and (4) both presume the salience of the Open command. The contexts in which (3) can be used are characterized by the question *What should we do by the Open command?*. It is also possible to use (3) in a context characterized by the question *What should we do?* 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 in particular does not have to be salient. (3) does not refer to a specific file.

(4), grammatical in Czech but ungrammatical in Russian and Bulgarian, presumes the Open command and a file to be salient, which is why we used the definite article in English. The appropriate contexts for (4) in Czech are characterized by the question *What should we do with the file by the Open command?*.

The above examples illustrate that differences in word order very often correspond to differences in information status of the entities and processes about which the text is, 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 word order variants.[3]

Next, let us consider the contextual appropriateness of the word order variants in (6-11) with respect to the context given in (5). In (6), we show the word orders appearing in the manual. (7-11) are the permutations of its main syntactic groups.

(5) Open the Multiline styles dialog box using one of the following methods.

(6)	**Cz**	Z	menu	Data vyberte	*Multiline Style.*
	Ru	V	menju	Data vyberite punkt	*Multiline Style*
	Bu	Ot	menjuto	Data izberete	*Multiline Style*
	gl	From|In	menu	Data choose	*Multiline Style*

(7)	**Cz**	Vyberte	z menu	Data	*Multiline Style.*
	Ru	Vyberite	v menju	Data punkt	*Multiline Style.*
	Bu	Izberete	ot menjuto	Data	*Multiline Style.*

[3] We assume a neutral intonation pattern in all examples in 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.

gl Choose from|in menu Data *Multiline Style*

(8) **Cz** Vyberte *Multiline Style* z menu Data.
Ru Vyberite punkt *Multiline Style* v menju Data.
Bu Izberete *Multiline Style* ot menjuto Data.
gl Choose *Multiline Style* from|in menu Data

(9) **Cz** *Multiline Style* vyberte z menu Data.
Ru Punkt *Multiline Style* vyberite v menju Data
Bu *Multiline Style* izberete ot menjuto Data.
gl *Multiline Style* choose from| menu Data

(10) **Cz** Z menu Data *Multiline Style* vyberte.
Ru ? V menju Data punkt *Multiline Style* vyberite
Bu ⋆ Ot menjuto Data *Multiline Style* izberete.
gl From|In menu Data *Multiline Style* choose

(11) **Cz** *Multiline Style* z menu Data vyberte.
Ru ? Punkt *Multiline Style* v menju Data vyberite.
Bu ⋆ *Multiline Style* ot menjuto Data izberete.
gl *Multiline Style* from|in menu Data choose

(6-11) all constitute grammatically well formed sentences in Czech as well as in Russian. In Bulgarian, (10) and (11) are ungrammatical. In Czech and Russian, it appears that both versions (7) and (8) could be used instead (6) in the context of (5); however, the remaining versions (9-11) could not. (9) does not fit into the context of (5) because it presumes salience of Multiline Style, which is not established in this context. (10) and (11) in Czech could only be used felicitously in a rather restricted type of contexts, namely those where the verb or the action referred to is to be interpreted as contrasting some other verb or action with the same participants and circumstances. In Russian, (10) and (11) sound strange, though they are grammatical.

It should be stressed that (10) and (11) where the verb appears in the sentence-final position are only unsuitable in the given context. There are contexts where such word order would be perfectly suitable; for instance, (12) provides such context for (13):

(12) Open a new file. Draw a schema.

(13) **Cz** A nyní soubor uložte.
Ru I teperj fajl sochranite.
gl And now file save.
En And now save the file.

The file is established in the context, and the important information conveyed by this sentence is the next action to be performed with it.

All the examples discussed above consist of instructions expressed in the imperative form, so there is no explicit Actor. In Slavonic languages, it is also possible to formulate instructions in indicative mood with active voice, with an explicit Actor corresponding to the user. Since our languages are all characterized as *pro-drop*, the Subject is dropped when the user is referred to by a personal pronoun (usually first person plural) in the Subject position, as in the Czech example in (14). The ordering possibilities are the same as in the imperative version in such a case.

(14) Otevřeme soubor příkazem Open.
We-open file-acc command-instr Open
One opens a file by the Open command.

When the Subject is non-pronominal, however, it undergoes the same word ordering mechanisms as other elements in the clause. We illustrate this by the set of Czech sentences with varying word order in (15-18).

(15) Příkaz *Open* otevírá soubor.
Command-nom *Open* opens file-acc.
The *Open* command opens a file.

(16) Soubor otevírá příkaz *Open*.
File-acc opens command-nom *Open*.
The file is opened by the *Open* command.

(17) Příkaz *Open* soubor otevírá.
Command-nom *Open* opens file-acc.
The *Open* command opens the file.

(18) Soubor příkaz *Open* otevírá.
File-acc command-nom *Open* opens.
The file is opened by the *Open* command.

The variants in (15) and (16) are the two more likely ones to be encountered. (15) does not presume the salience of a file, while (16) does. Both (17) and (18) presume that both the Open command and the file are salient. The difference between them is that in the latter, the file appears to be contrasted with some other entity.

Last but not least, Slavonic languages can express instructions by sentences in indicative mood in passive voice, using a reflexive construction, as illustrated by the Czech examples in (19-22).

(19) Příkazem *Open* se otevírá soubor.
Command-instr *Open* refl opens file-nom.
A file is opened by the *Open* command.

(20) Soubor se otevírá příkazem *Open*.
File-nom refl opens command-instr *Open*.
The file is opened by the *Open* command.

(21) Příkazem *Open* se soubor otevírá.
Command-instr *Open* refl opens file-nom.
By the *Open* command, the file is opened.

(22) Soubor se příkazem *Open* otevírá.
File-nom refl command-instr *Open* opens.
The file is opened by the *Open* command.

In Russian, reflexive forms are created with suffix '-sja' attached to the main verb, and therefore no particular issue of word order arises. In Czech as well in Bulgarian, the reflexive particle 'se' is a clitic, and is therefore subject to the special ordering constrains concerning clitics: they have to be placed in the so-called *Wackernagel's position* (see Section 4).

To summarize, sentences which differ only in word order (and not in syntactic realizations of constituents) are not freely interchangable in a given context. The comparison of word order variation in Czech, Russian and Bulgarian also illustrates that the degree of word order freedom is not the same across the three languages under consideration. Bulgarian is a language with a more restricted word order than Czech and Russian. And also among the latter two, some small differences have been identified when analyzing simple examples.

One conjecture is that whereas Czech and Russian can form grammatical sentences with more than one constituent expressing an experiential element preceding the verb, this seems impossible in Bulgarian. Like English, Bulgarian appears to allow for at most one experiential element preceding the verb. On the other hand, we found out that Bulgarian allows more word order freedom than English with respect to the elements following the verb, at least in some configurations. Our conjecture is that the Bulgarian word order follows information structure in this case. In Czech and Russian, as expected, there is even more ordering flexibility. In most cases information structure can be reflected simply be the appropriate ordering of elements regardless of the grammatical structure of the sentence. However, we are aware of some cases, especially in Russian, where a simple permutation of elements does not yield a suitable ordering. In particular, we encountered a problem with an attempt to order nominalization expressing Means at the beginning of a sentence. In some cases Russian also seems to prefer a passive construction with reflexive form while Czech uses a permutation in active voice (cf. [1] for a detailed discussion).

In order to handle word order in Slavonic languages, and in particular within the AGILE project, we need to be able to capture not only the structural restrictions specific for the individual languages, but also the influence of the information status of the entities being referred to on the ordering of the expressions in the generated sentences.

2.2 Word Order and Information Structure

In the Slavonic linguistic tradition, the main factor determining the linear ordering within a sentence is considered to be the *information structure*. We are using the term information structure 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". Within the Czech linguistic tradition ensuing from the Prague School, information structure is referred to as *topic-focus articulation* (see [15] for a concise account), or *functional sentence perspective* [5]. Another terminology, where information structure is called *information packaging*, was introduced by Chafe in the 70's and used by Vallduví [16]. Halliday [8, 9] makes a distinction between the *thematic structure* of the sentence and its *information structure*.

In every approach to information structure, the clause is considered to consist of (at least) two parts. The often used oppositions are, e.g. *Theme-Rheme* [14, 9, 4], *Topic-Focus* [15], *Background-Focus* [10], *Ground-Focus* [16]. Sometimes, the authors introduce further sub-divisions.[4]

In order to account for the word ordering phenomena within the AGILE project, we build on the insights the SFG and FGD frameworks. As a prelude to the presentation of our approach, we now briefly overview the essential SFG and FGD notions relevant for word ordering.

Thematic and Information Structure in SFG Some of the considerations related to word order have been reflected in the SFG framework, which we take as the starting point for developing the linguistic specifications in the AGILE project. According to [9], a clause as a message consists of a Theme combined with a Rheme, and in this configuration, the Theme is the ground from which the clause is taking off. As noted earlier, Halliday distinguishes between the thematic structure of a clause and the information structure. The latter is the distinction between Given and New within an information unit: the speaker presents information to the listener as recoverable (Given) or not recoverable (New). The thematic structure and information structure are closely related but not the same. Whereas the Theme is what the experiential items the speaker chooses to take as the point of departure, the Given is what the speaker believes the listener already knows or has accessible.

The notion of Theme tells us a number of things about "the first" position in the clause, but it does not tell us much about the word order of "the rest" of the clause. Presumably, Halliday leaves this to be decided by the grammatical structure. However, in languages with a high degree of free word order the grammar is not very strict about the placement of the groups and phrases within the clause. The examples we discussed above showed that ordering in our languages is to a great extent determined by what is presumed to be salient in the context. This means ordering depends on information structure. These issues have been studied in detail in the Praguian FGD framework [15]. We incorporate the most essential ideas into the AGILE account of word order in Slavonic languages.

Topic-Focus Articulation in FGD FGD works with a notion of information structure consisting of one dichotomy, called *topic-focus articulation* (TFA). TFA is defined on the basis of a distinction between contextually bound (CB) and non-bound (NB) items in a sentence [15] (or cf. [11] for an overview). The motivation behind this distinction corresponds to that underlying the

[4] Within the scope and aim of the present paper, we refrain from a more in- depth discussion of the differences between the various approaches to information structure (but see [7, 11]).

Given/New dichotomy in SFG. 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 (see [6] for a summary).

The ordering of NB items in a sentence follows the so-called *systemic ordering* (SO). SO is a language specific ordering of complementations, i.e. "arguments" and "adjuncts", of verbs, nouns, adjectives or adverbs which corresponds to neutral word order. It may differ from one language to another, but is considered universal within a given language. SO in Czech has been studied in detail (see [15]). The SOs of Russian and Bulgarian have not yet been studied in general. We expect 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 [1].

We summarize the FGD claims concerning word order below.

1. The main principle of word order in Czech is that the Topic precedes the Focus. Since the Topic may be empty (esp. in discourse-initial sentences) or may be deleted on the surface due to ellipsis, it is possible that the surface form of some sentences only consists of the realisations of elements belonging to the Focus.

2. In the primary cases when the Topic consists of the CB elements, and the Focus of the NB ones, one can say that the CB elements precede the NB elements.
   A more general formulation of this principle uses the degrees of the so-called *communicative dynamism* (CD, [15]): in primary cases, CD and the surface word order correspond to each other quite closely, at least within clauses. So, the ordering from left to right in the surface realisation corresponds to the increasing degrees of CD. There are the following exceptions to this principle in Czech:

   (a) clitics: they have to be placed in the so-called Wackernagel's position, characterized roughly as the position between the first and the second element in a clause.[5]

   (b) the main verb: its preferred default (unmarked) placement is after, but not necessarily immediately following, the surface Subject, if there is one

In the next section we show how the SFG and FGD ideas concerning information structure can be used in an integrated approach to generate contextually appropriate ordering of clause elements.

[5]Naturally, this leaves to be defined what "element" means. It is easy to show that "first element" does not equate to "first constituent", since the element can be of arbitrary complexity. We try to use the notion of Theme for this purpose (see Section 3.1).

3 The Approach in AGILE

As the starting point for specifying the principles of word ordering in the context of AGILE, we combine this FGD-based strategy which reflects information structure with the possibility of thematization in the usual SFG spirit. For Czech and Russian, we need to allow for more freedom in word order (i.e., a looser relation between ordering and grammatical structure) than in Bulgarian (cf. the summary in Section 2.1).

We propose to preserve the SFG notion of Theme as being the first experiential element in the clause. This conception appears useful in any language in order 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 it is not needed to look for a motivation for a particular ordering based on information structure.

For ordering of non-thematic constituents within a clause, for which their order 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.

Contextual boundness is treated as a local feature, i.e., a complex contextually bound constituent can contain "locally" contextually non-bound constituents, and vice versa. For instance, a complex sentence can contain one CB and one NB clause, and within each of them, elements 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).

A specific point concerns the placement of clitics in Slavonic languages. In the sentences we are have been generating in AGILE so far, it appears sufficient to place clitics after the Theme.

3.1 Word Ordering Algorithm

To summarize our approach, we present a word ordering algorithm. The algorithm is presented in abstract form in Figure 1. Using $<$ for linear precedence, it can be schematized as follows:

$$\texttt{Theme} < \texttt{Clitics} < \texttt{Rest-CB} < \texttt{Verb} < \texttt{Rest-NB}$$

The Theme is determined by text organization. If no element is explicitly chosen as Theme, the thematic position is filled by the first CB element. Any clitics are placed after the Theme. Their mutual order is determined by the grammar. For ordering of the non-thematic constituents within a clause, for which their order is not determined by the syntactic structure, we use systemic ordering in combination with the CB/NB distinction. The NB elements are ordered by SO. The ordering of the CB elements can be (i) specified on the basis of the context, (ii) restricted by the grammatical structure, (iii) follow SO. The verb is placed between the last CB and the first NB element, unless it is itself the Theme.

Figure 1: Abstract algorithm for word order

```
Given: a list G of ordering constraints imposed by the grammar,
       a list L1 of constituents that need to be ordered,
       a list Delta giving ordering of CB constituents,

    create empty lists LC and LN % LC for CB items, 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.
    if the verb is CB,
        then    Order the verb at the end of LC
                Order the remainder according to D
        else    Order all elements in LC according to D

    % thus, if e precedes f in D, then e precedes f in LC except for the verb.

    if G is not empty then
        Order elements in L1 using ordering constraints in G
```

The proposed ordering algorithm as such is the same for all the three languages under consideration. What may differ is first of all the systemic ordering, and therefore the default ordering of non-bound elements. Further differences between the languages are encountered in the constraints on which elements can be ordered rather freely and which elements are subject to ordering requirements posed by the syntactic structure.

3.2 Text Generation Overview

Within AGILE, we aim at so-called "end-to-end" generation: A user specifies an A-box using an authoring interface, and obtains as output a text realizing that A-box, in a particular text style.

The A-box serves as input to the Text Structuring Module (TSM) which yields a set of formulas in a Sentence Planning Language (SPL) for the sentences to be generated to convey the given content, in a particular text style. To guide the text planning done by the TSM, we employ text structure elements that correspond to identifiable parts in an A-box configuration, and text templates that specify particular text styles.

The TSM is described in detail in [13]. The essential ideas can be summarized as follows. The major component is formed by the systemic networks for text structuring. In the spirit of [3], we construct a region that defines an additional level of linguistic resources for the level of *genre*. The region enables the composition of text structures in a way that is very similar to the way the lexico-grammar builds up grammatical structures. In fact, by using KPML to implement the means for text structuring, it easily facilitates smooth interaction between global level text generation (strategic generation) and lexico-grammatical expression (tactical generation)[6]. The region follows out the viewpoint that text templates and text structure elements are essentially orthogonal ideas. Therefore it consists of two parts. One part deals with interpreting the A-box in terms of text structure elements. By traversing the network that the systems of this part make up, we obtain a text structure for the A-box conforming to the way the A-box structures the content. The other part of the region imposes constraints on the realisation of the text structure elements that are being introduced by traversing the other part of the region. Naturally it will depend on our choice of a particular text template (style) which constraints will be imposed. These choices are made through interaction between the user and the system.

After a text structure is made by traversing the systemic networks, we divide the A-box into smaller A-boxes that can be associated with the text structure elements from which the text structure is composed. In the setting of KPML this can easily be done using so-called ID-inquiries. Finally, SPLs are generated, using these individuated A-boxes and the realization constraints imposed by text templates.

The SPLs then serve as input to language-specific tactical generators that actually generate these sentences. 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 a smooth flow of the text can be assured.

The idea that we implement is that the SPLs contain information about contextual boundness vs. non-boundness of each clause element. When the corresponding syntactic constituents are generated, the ordering algorithm discussed in Section 3.1 applies. So for instance, once all the constituents of a clause have been created, their ordering is determined. As we stated above, grammatical constraints on the ordering are taken into account first. The ordering according to information structure, i.e. contextual boundness vs. non-boundness, is applied next. Finally, if the ordering of some elements is still undecided, defaults are applied. Such ordering process takes place at every level of grammatical structure. Thus, what is being ordered are (i) clauses in a complex sentence, (ii) elements of individual simplex clauses, i.e. prepositional phrases, nominal groups, adverbs, (iii) elements within each nominal group, etc.

4 Concluding Remarks

We have presented an approach to handling word order in Czech, Russian and Bulgarian in the course of automatic natural language generation from an underlying knowledge representation. We have demonstrated that the information that is expressed in one language by a property of a nominal expression, such as definitness, might find a different expression in other languages. In particular,

[6]In this way we try to overcome the notorious problem known as the *generation gap* in which a text planning module lacks control over the fine-grained distinctions that are available in the grammar. In our case, both text planning and sentence planning are integrated into one and the same system (albeith organized in a stratificational manner).

in Slavonic languages, which are characteristic by a relatively high degree of free word order, ordering can reflect familiarity of entities with respect to the context.

We believe that some general issues can be raised on the basis of the discussion we have provided. With respect to the focus of the present workshop on generation of nominal expressions, there is the question of what should be considered as the task of nominal expression generation. Clearly, the preceding context does not only determine internal properties of nominal expressions, but also their mutual ordering within each clause.

Consequently, word ordering gives rise to system modularity and architecture questions that need to be addressed for nominal expressions as well as for other components of the generation process.

The current approach has the following shortcoming. It handles ordering after the determination of syntactic structure and after the completion of syntactic constituents at a given level, i.e. the clause. Word order is currently the only reflection of information structure the sentences we generate. Such approach does not enable information structure to influence the grammatical structure and the features of the constituents that are generated.

References

[1] Elena Adonova, John Bateman, Nevena Gromova, Anthony Hartley, Geert-Jan Kruijff, Ivana Kruijff-Korbayová, Serge Sharoff, Hana Skoumalová, Lena Sokolova, Kamenka Staykova, and Elke Teich. Formal specification of extended grammar models. AGILE project deliverable, University of Brighton, UK, `http://ufal.mff.cuni.cz/~agile/reports.html`, 1999.

[2] J. A. Bateman, R. Kasper, J. D. Moore, and R. Whitney. A general organization of knowledge for natural language processing: the Penman Upper Model. Technical report, USC/ISI, Marina del Rey, California, 1990.

[3] John A. Bateman. Enabling technology for multilingual natural language generation: the KPML environment. *Natural Language Engineering*, 1(1), 1997.

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

[5] Jan Firbas. *Functional Sentence Perspective in Written and Spoken Communication*. Studies in English Language. Cambridge University Press, Cambridge, 1992.

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

[7] Eva Hajičová and Ivana Kruijff-Korbayová. On the notion of topic. In *Prague Linguistic Circle Papers*. John Benjamins Publishers, Amsterdam, the Netherlands, in print.

[8] M.A.K. Halliday. Language structure and language function. In John Lyons, editor, *New Horizons in Linguistics*, pages 140–165. Penguin Books Ltd., 1970.

[9] M.A.K. Halliday. *An Introduction to Functional Grammar.* Edward Arnold, London, 1985.

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

[11] Ivana Kruijff-Korbayová. *The Dynamic Potential of Topic and Focus: A Praguian Discourse Representation Theory.* Ph.d. thesis, Charles University, Prague, Czech Republic, `http://kwetal.ms.mff.cuni.cz/~korbay/publications.html`, 1998.

[12] Ivana Kruijff-Korbayová. Generation of instructions in a multilingual environment. In *Proceedings of the Conference on Text, Speech and Dialogue (TSD'98), Brno, Czech Republic, September 1998*, pages 67–72, 1998.

[13] Ivana Kruijff-Korbayová and Geert-Jan Kruijff. Text structuring in a multilingual system for generation of instructions. In *Proceedings of the Conference on Text, Speech and Dialogue (TSD'99), Czech Republic*, 1999, to appear with Springer-Verlag.

[14] Petr Sgall, Eva Hajičová, and Eva Benešsová. *Topic, focus and generative semantics.* Kronberg, Taunus, 1973.

[15] Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The meaning of the sentence in its semantic and pragmatic aspects.* Reidel, Dordrecht, The Netherlands, 1986.

[16] Enric Vallduví. *The Informational Component.* Garland, New York, 1992.