Handling Word Order in a Multilingual System for Generation of Instructions

Ivana Kruijff-Korbayová and Geert-Jan Kruijff

Institute of Formal and Applied Linguistics, Charles University, Czech Republic,
{korbay,gj}@ufal.mff.cuni.cz,
WWW home page: http://ufal.mff.cuni.cz

Abstract. Slavic languages are characteristic by their relatively high degree of word order freedom. In the process of automatic generation from an underlying representation of the content, we have to ensure that a semantically and contextually appropriate word order is chosen. In this paper, we elucidate *information structure* as the main factor determining word order in Slavic languages, and we present an approach to handling word order in text generation in the context of the AGILE project [4].[1]

1 Introduction

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 multidimensional. 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 in the language system in general can be discerned 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.

[1] AGILE (Automatic Generation of Instructions in Languages of Eastern Europe) is supported by the European Commission within the COPERNICUS programme, grant No. PL961104. We would like to thank our colleagues from the Charles University, the University of Brighton, the Saarland University, the Bulgarian Academy of Sciences and the Russian Research Institute of Artificial Intelligence.

Information structure is considered the main factor determining the linear ordering within a sentence in the Slavonic linguistic tradition. It is considered relevant for both word order and clause order. Hereby, 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, information structure has been referred to as "aktuální členění", functional sentence perspective and topic-focus articulation. Another terminology was introduced by Chafe in the 70's, where information structure is called *information packaging*. Halliday 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*, *Topic-Focus*, *Background-Focus*, *Ground-Focus*. Sometimes, the authors introduce further sub-divisions. For a discussion of some of the differences between various approaches see [5].[2]

In the AGILE project, we cannot consider all these different approaches to information structure. In the end, we have to work with just one approach for the sake of our linguistic specifications. Currently we restrict ourselves to two approaches: (i) Halliday's thematic structure [3] as developed in the Systemic Functional Grammar framework (SFG) is chosen because our grammars are based on the SFG framework; (ii) the topic-focus articulation approach developed in Prague within the framework of Functional Generative Description (FGD, [8]) 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.[3]

In this paper, we concentrate on the issue of controlling word order in the course of automatic natural language generation. We begin by an illustration of the differences in semantic and contextual appropriateness of sentences with varying word order. Then we present the main principles which enable us to determine a suitable word order when generating a sentence conveying a particular content in a given language in a particular context. Finally, we present our word ordering algorithm, based on combining the SFG and FGD insights, which applies to the constituents of a clause.

[2] For a detailed list of references to various approaches to information structure see [1].

[3] Due to the restricted size of this paper, we cannot overview and compare the SFG and FGD approaches in detail, so the interested reader should consult [1].

2 Word Order Variation

We first present some word order variations in the three languages using an example from the set of texts generated in AGILE, which are adapted from a CAD/CAM system user guide. Let us consider the context given in (1).

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

In (2) we show the following sentence from the manual in its original word order. In (3) through (7) are the permutations of its main syntactic groups:

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

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

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

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

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

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

(2) - (7) all constitute grammatically well formed sentences in Czech as well as in Russian. In Bulgarian, (7) and (6) are ungrammatical. It is thus apparent that for sentences with a moderate number of syntactic groups there are usually quite a few grammatically well formed word order variants. However, one should not interpret the high degree of freedom in word order as arbitrariness.

In Czech and Russian, it appears that both versions (3) and (4) could be used instead (2) in the context of (1); however, the remaining versions (5) - (7) could

not. (5) does not fit into the context of (1) because it presupposes contextual familiarity of *Multiline Style*, which it is not in this context. (6) and (7) 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, (6) and (7) sound strange, though they are grammatical.

Thus, sentences which differ only in word order (and not in syntactic realizations of constituents) are not freely interchangable in a given context (cf. [1, 6] for more discussion). The comparison also illustrates that the degree of word order freedom is not the same in across the three languages under consideration. 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 in every language.

3 Information Structure and Word Order

According to [3], 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 [8]. We incorporate the most essential ideas into the AGILE account of word order in Slavonic languages.

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. The motivation behind this distinction corresponds to that underlying the 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 [2] for a summarizing discussion).

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 [8]). We expect the SO for the main types of complementations in Russian and Bulgarian is similar to the Czech one, but there can be differences [1].

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. Namely, in Bulgarian there seems to be the restriction that only one experiential element may precede the verb. Other restrictions or preferences can be included in the grammars of the specific languages.

A specific point concerns the placement of clitics in Slavic languages: they must appear in what is called the Wackernagel position. In the sentences we are generating in AGILE, it appears useful to use the Theme to identify the position for clitics.

4 Word Ordering Algorithm

The algorithm is presented in abstract form in Figure 1. Using ^ for linear precedence, it can be schematized as follows:

```
Theme^Clitics^Rest-CB^Verb^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.

The proposed ordering algorithm is the same for all the three languages under consideration. What differs across the languages are constraints on which elements can be ordered rather freely in accordance to information structure, and which ones are subject to ordering requirements posed by the syntactic structure.

The current approach is satisfactory to the extent that is it satisfactory to consider word order as a "second order" phenomenon, i.e. as an ordering that is applied to constituents in a structure that has already been generated. Such approach does not enable information structure to influence the grammatical structure of the constituents that are generated.

Fig. 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
```

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. Eva Hajičová. *Issues of sentence structure and discourse patterns*, volume 2 of *Theoretical and computational linguistics*. Charles University, Prague, Czech Republic, 1993.
3. M.A.K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold, London, 1985.
4. 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.
5. Ivana Kruijff-Korbayová and Eva Hajičová. Topics and centers: a comparison of the salience-based approach and the centering theory. *Prague Bulletin of Mathematical Linguistics*, (67):25–50, 1997. Charles University, Prague, Czech Republic.
6. Ivana Kruijff-Korbayová and Geert-Jan Kruijff. Contextually appropriate ordering of nominal expressions. In *Proceedings of the ESSLLI99 workshop "Generating Nominals", August 1999*, 1999.
7. 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, September 1999*, 1999.
8. Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The meaning of the sentence in its semantic and pragmatic aspects*. Reidel, Dordrecht, The Netherlands, 1986.