Linear order as higher-level decision: Information Structure in strategic and tactical generation

Geert-Jan M. Kruijff,
Ivana Kruijff-Korbayová
Computational Linguistics
University of the Saarland
Saarbrücken, Germany
⟨{gj,korbay}@coli.uni-sb.de⟩

John Bateman
Applied Linguistics
University of Bremen
Bremen, Germany
⟨bateman@uni-bremen.de⟩

Elke Teich
Applied Linguistics
University of the Saarland
Saarbrücken, Germany
⟨E.Teich@mx.uni-saarland.de⟩

Abstract

We propose a multilingual approach to characterizing word order at the clause level as a means to realize information structure. We illustrate the problem with three languages which differ in the degree of word order freedom they exhibit: Czech, a free word order language in which word order variation is pragmatically determined; English, a fixed word order language in which word order is primarily grammatically determined; and German, a language which is between Czech and English on the scale of word order freedom. Our work is theoretically rooted in previous work on information structuring and word order in the Prague School framework as well as on the systemic-functional notion of Theme. The approach we present has been implemented in KPML.

1 Introduction

The aim of this paper is to describe an architecture that addresses how information structure can be integrated into strategic and tactical generation. We focus primarily here on the tactical aspect of how word order (henceforth: WO) may function as a means of realizing information structure. The approach we take is multilingually applicable. It is implemented in KPML (Bateman, 1997b; Bateman, 1997a) and has been tested for Czech, Bulgarian and Russian as three Slavonic languages with different WO properties, as well as for English. The algorithm itself is not KPML-specific: it combines the idea of WO constraints posed by the grammar, with a complementary mechanism of default ordering based on information structure. The algorithm could thus be applied in other systems wich allow multiple sources of ordering constraints.

Information structure is a means that a speaker employs to indicate that some parts of a sentence meaning are context-dependent (“given”), and that others are context-affecting (“new”). Information structure is therefore an inherent aspect of sentence meaning, and it contributes in an important way to the overall coherence of a text. While it is commonly accepted that information structuring is a major source of constraints for the organization of a given content in a particular linear order in many languages, there is very little work in Natural Language Generation that explicitly models this relation.

From a practical perspective, in the most commonly employed generation systems such as KPML, FUF (Elhadad, 1993; Elhadad and Robin, 1997) or REALPRO (Lavoie and Rambow, 1997), linear ordering comes as a by-product of other grammatical choices. This is fine for tactical generation components and it is sufficient for languages with grammatically determined WO (‘fixed’ WO languages), such as English or Chinese. However, most languages have some WO variability and this variation usually reflects information structure. When languages in which linear order is *primarily* pragmatically determined are involved, such as the Slavonic languages we have dealt with, a number of problems become immediately apparent.

A comprehensive account of WO variation for natural language generation that is reusable across languages is thus required. Such an account needs to represent linearization as an explicit decision-making process that involves *both* the representation of the language-specific linear ordering possibilites *and* the representation of the language-specific (and possibly cross-linguistically valid) motivations for particular linearizations. Again, while the former is catered for in most tactical generation systems, only selected aspects of the latter have been dealt with and only for selected languages (e.g., (Hoffman, 1994; Hoffman, 1995; Hakkani et al., 1996)).

For example, (Hoffman, 1994) proposes a treatment of WO in Turkish using a categorial grammar framework (CCG, (Steedman, 2000)) and relating this to Steedman's (earlier) account of information structure (Steedman, 1991). However, the most important issue, that of providing an integrated account of how information structure guides the choice of (or, is realized by) linear ordering, is left unsolved (Kruijff, 2001).

Given that in many languages, information structure is the major driving force for WO variation, it is indeed the most straighforward idea to couple an account of information structure with the choice of linear ordering. However, for multilingual application, the particular challenge is to develop a solution that can be applied, no matter at which point on the free-to-fixed WO cline a language is located.

The approach to WO proposed in this paper is a move in exactly this direction. We start in §2 with presenting data from Czech, German and English that motivate the perspective we take on information structure, and its role in generating coherent discourse. In §3 we introduce the linguistic notions employed in the present account. In §4 we discuss how information structure fits into a general system architecture, and we discuss the implementation of the strategic generation component on the basis of KPML. We continue with an elaboration of the role of information structure in tactical generation, presenting an algorithm for generating contextually appropriate linearization, given a sentence's information structure, and illustrate its implementation on Czech and English examples (§5). We conclude the paper with a summary (§6).

2 Linguistic motivation

There are a number of factors commonly acknowledged to 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*. Cross-linguistically, these factors may be involved in constraining linear ordering to varying degrees. English, for instance, is an example of a language in which WO is rather rigid, i.e., 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 particular types of grammatical constructions, such as clefting and pseudo-clefting, or definiteness/indefiniteness of the nominal group. Czech, in contrast, which has a rich case system and no definite or indefinite article, belongs to the so-called "free word order" languages, where the same effects are achieved by varying WO. Finally, German lies between English and Czech in the spectrum between fixed and free WO. We illustrate the general point that WO selections are related to information structure by appropriateness judgements of some examples of instructions in Czech, German and English.[1]

(1)	Otevřeme	příkazem	`Open`	soubor.
	open-1PL	command-INS	`Open`	file-ACC

Sie	öffnen	eine	Datei	mit	dem	Befehl	`Open`.
You	open	a	file	with	the	command	`Open`.

Open a file with the `Open` command.

The ordering in (1) is neutral in that no particular contextual constraints hold with respect to the newsworthiness of any of the elements expressed in this clause. This kind of ordering can

[1] The English examples use imperative mood, while the Czech and the German examples use indicative mood as the most common way of conveying instructions of the discussed type. Alternatively, both Czech and German can use also imperatives or infinitives for instructions, but these are considered less polite than the indicative versions. Last but not least, instructions can also be formulated in indicative mood with passive voice in both Czech and German.

be elicited by the question *What should we do?*.[2] We follow Prague School accounts (Firbas, 1992; Sgall et al., 1986) in calling this neutral ordering the *systemic ordering* (cf. also §5). Alternatively, (1) 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.

(2) Otevřeme soubor příkazem Open.
open-1PL file-ACC command-INS Open

Sie öffnen die Datei mit dem Befehl Open.
you open the file with the command Open.

"Open the file with the Open command."

(3) Soubor otevřeme příkazem Open.
file-ACC open-1PL command-INS Open

Die Datei öffnen Sie mit dem Befehl Open.
the file open you with the command Open.

"Open the file with the Open command."

The word order variants illustrated in (2) and (3) are appropriate when some file is active in the context (Chafe, 1976), for instance when the user is working with a file. In (2), the action of opening is also active; in (3) it can, but does not have to be active, too. The contexts in which (2) and (3) 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 (2), example (3) can be used if file is contrasted with another entity. In German, this contrast is required, whereas in Czech it is optional. In English, intonation could mark whether contrast is required.

(4) Příkazem Open otevřeme soubor.
command-INS Open open-1PL file-ACC

Mit dem Befehl Open öffnen Sie eine Datei.
with the command Open open you a file.

With the Open command, open a file.

(5) Příkazem Open soubor otevřeme.
command-INS Open fileACC open-1PL

Mit dem Befehl Open öffnen Sie die Datei.
with the command Open open you a file.

With the Open command, open the file.

The contexts in which (4) can be used are characterized by *What should we do with the* Open *command?*. While (4) does not refer to a specific file, in (5) an activated file is presumed. (5) is appropriate in contexts characterized by *What should we do to the file with the* Open *command?*.

It is also possible to use (4) in a context characterized by *What should we do?*, and (5) 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 activated.

(6) Soubor příkazem Open otevřete.
file-ACC command-INS Open open-I2PL

Die Datei öffnen Sie mit dem Befehl Open.
the file open you with the command Open

Open the file with the Open command.

Example (6) is like (5) in that it is appropriate when both a file and the Open command are activated. The contexts in which (6) can be appropriately used can be characterized by *What should we do to the file with the* Open *command?*. Unlike (5), (6) can also be used when file is contrasted with another entity. In German, there is no difference in word order between (6) and (3) (they differ only in intonation). This is a result of the strong ordering constraint in German to place the finite verb as second (in independent, declarative clauses). In Czech verb secondness also plays a role, but it is much weaker.

Analogous judgements concerning contextual appropriateness apply to WO variants in different mood and/or voice (when available in the individual languages). The orders in which the verb is first do not presume the activation of either a file or a command. The orders in which 'file' precedes the verb appear to presume an active file, the orders in which 'command' precedes the verb appear to presume the activation of a command. When both 'file' and 'command' precede the verb, the activation of both a file and a command appears to be presumed.

These judgements show that differences in WO (in languages with a more flexible WO then English, e.g., Czech and German) very often correspond to differences in how the speaker presents

[2] We use questions for presentational purposes to indicate which contexts would be appropriate for uttering sentences with particular WO variants. Such question-answer pairs are known as **question tests** (Sgall et al., 1986).

the information status of the entities and processes that are referred to in a text, in particular, whether they are assumed to be already familiar or not, and whether they are assumed to be activated in the context. Note that in English, the same distinction is expressed by the use of a definite vs. an indefinite nominal expression, i.e. ‘a|the file’.

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

3 Underlying notions

In order to provide constraints for WO decisions within our generation architecture, we require mechanisms through which particular patterns of information structuring can constrain 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; these approaches are briefly introduced below.

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; Halliday, 1985) distinguishes between the *thematic structure* of a clause and its *information structure*: Whereas the *Theme* is “the starting point for the message, it is the ground from which the clause is taking off” (Halliday, 1985, 38), information structure concerns the distinction between the *Given* as “what is presented as being already known to the listener” (Halliday, 1985, 59), and the *New* as “what the listener is being invited to attend to as new, or unexpected, or important” (*ibid*).

3.1 Information structure and ordering

In Halliday’s original approach (Halliday, 1967), the basic assumption for English and also for other languages is that ordering, apart from being grammatically constrained, is iconic with respect to “newsworthiness”. So on a scale from Given to New information, the “newer” elements would come towards the end of the information unit, the “newest” element bearing the nuclear stress. This approach relies on the possibility of giving a complete ordering of all clause elements with respect to their newsworthiness.

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)). Also from the viewpoint of CD, 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 the *language specific* canonical ordering of clause elements (complements and adjuncts), as well as of elements of lower syntactic levels, with respect to one another.

For the current purposes we concentrate on the SO for a subset of the clause elements that are discerned in FGD. We use the following SOs for the Slavonic languages and for English and German:[3]

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

SO for English: Actor < Addressee < Patient < SpaceLocative < TemporalLocative < Means < Source < Destination < Purposedependent

SO for German: Actor < TemporalLocative < SpaceLocative < Means < Addressee < Patient < Source < Destination < Purpose

[3] The labels we use for the various types of elements are a mixture of FGD and SFG terminology.

The SO for the Slavonic languages is based on the one for Czech (Sgall et al., 1986); the only difference is that we have placed Patient before Source ('from where'). 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)). 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 5.2). The SO for German is based on (Heidolph et al., 1981, p.704).

The informational status of elements is established through deviation of CD from the SO. 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 on the preceding discourse context. It may refer to an entity already explicitly referred to in the discourse, or an "implicitly evoked" entity. 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 can be used to constrain WO (at the clause level) as follows:

- The CB elements (if there are any) typically precede the NB elements.
- 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 SO (with the exceptions required by grammatically constrained ordering as described below). The default for communicative dynamism is SO.
- 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).

It is the above abstract ordering principles that underly the algorithm we present in §5.

3.2 Thematic structure

In all languages we looked at so far, 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, the latter can be readily addressed using the notion of Theme. For illustration, consider (8) in Czech, German and English, appearing in a context where it is preceded only by (7).

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

(8) Z menu `Data` vybereme `Style`.
From `Data` menu choose$_{1pl}$ `Style`.

Im Menü `Data` wählen Sie `Style`.
In menu `Data` choose you `Style`.

In the `Data` menu, choose `Style`.

The preceding context does not refer to the '`Data` menu' or make it active in any way. Working only with the notion of information structure discerning CB (Given) and NB (New) elements, one is thus unable to explain this ordering. On the other hand, the notion of *thematic structure* as a reflection of a global text organization strategy makes such explanation possible. In Halliday's approach, *Theme* has a particular textual function, that of signposting the intended development or "scaffolding" that a writer employs for structuring an extended text. In software instruction manuals, for example, we encounter regular thematization of (i) the location where actions are performed, (ii) the particular action that the user is instructed to perform, or (iii) the goal that the user wants to achieve (cf. (Kruijff-Korbayová et al., in prep) for a more detailed discussion).

4 Information structure and strategic planning

In this section we briefly describe how we integrate information structure into strategic generation, i.e. text- and sentence-planning. The

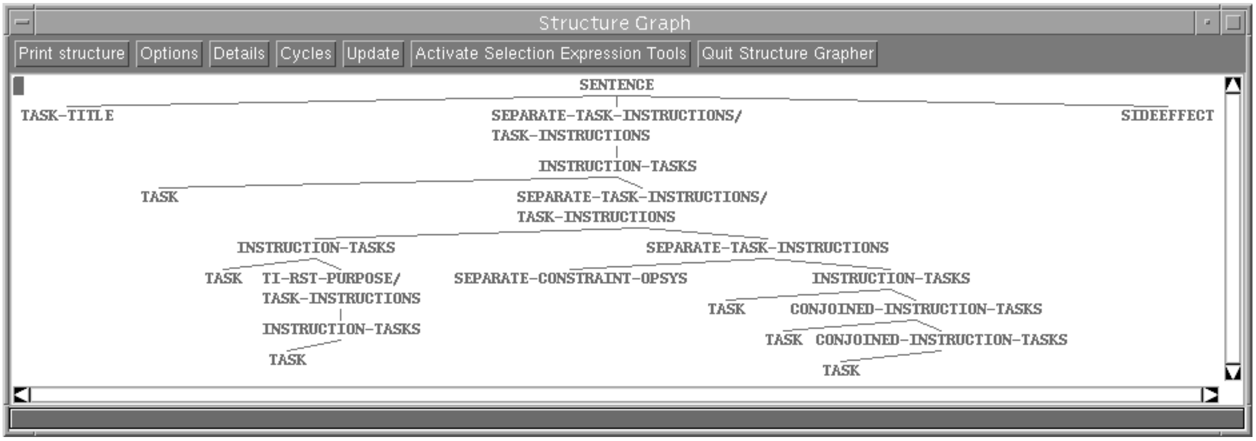

Figure 1: A text plan. In our system, a text plan organizes content into a linear fashion, showing where (and how) content might be aggregated syntactically (e.g. conjunction) or discursively (e.g. RST-relations). In the example above, the text plan specifies a text consisting of an overall goal (the title) and five substeps to resolve that goal (the tasks). The first task is a simple one, the second task is a complex formed around an RST-purpose relation, after which follows a conjunction of tasks. (The CONJOINED-INSTRUCTION-TASKS nodes indicate that the left-daughter node (a task) and the task dominated by the immediate non-terminal node above a CONJOINED-INSTRUCTION-TASKS node, are to be related by a conjunction.) The content to be realized is identified by the leaves of the text plan. Whenever a leaf is introduced in the text plan, the discourse model is updated with the content's (A-box) concepts. The sentence planner decends through the text plan depth-first. Thereby it gathers the leaves' content into sentence-specifications, following any indications of aggregation. It makes use of the discourse model to specify whether content should be realized as contextually bound (or not).

principle idea is that during text-planning, a discourse model is built that is then used in sentence-planning to determine a sentence's information structure.

We have developed a system using KPML. In KPML, generation resources are divided into interacting modules called *regions*. For the purpose of text-planning we have constructed a region that defines an additional level of linguistic resources for the level of *genre*. The region facilitates the composition of text structures in a way that is very similar to the way the lexico-grammar builds up grammatical structures. This enables us to have a close interaction between global level text generation and lexico-grammatical expression, with the possibility to accommodate and propagate constraints on output realization. While constructing a text plan, the text planner constructs a (rudimentary) discourse model that keeps track of the discourse entities introduced.

Text planning results in a text plan and a discourse model that serve as input to the sentence planner. The text plan is a hierarchical structure, organizing the content into a more linear fashion (see Figure 3.2). The sentence planner creates the input to the tactical generation phase as formulas of the Sentence planning Language (SPL, (Kasper, 1989)). The SPL formulas express the bits of content identified by the text plan's leaves, and can also group one or more leaves together (aggregation) depending on decisions taken by the text planner concerning discourse relations. Most importantly, during this phase of planning what content is to be realized by a sentence, the underlying information structure of that content is determined: 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 contextually bound (note that we are hereby making a simplifying assumption that in the current version of the sentence planner we equate contextual boundness with previous mention).

The text planner can also 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 some element as the default Theme. This can be the Subject (as in English), the least communicatively dynamic element (as in Czech); the choice of the default Theme in German is freer than in English, but more restricted than in Czech (cf. (Steiner and Ramm, 1995) for a discussion). 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 is also resolved by the grammar.

5 Realizing information structure through linearization

It is in the setting described in §4 that the issue of generating contextually appropriate sentences really arises. In this section we describe the word ordering algorithm (§5.1) and its application to Czech and English (§5.2).

5.1 Flexible word order algorithm

As discussed, constraints from various sources need to be combined in order to determine grammatically well-formed and contextually appropriate WO. Contextual boundness is used to constrain WO at the clause level as specified above. 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;

- information structure can determine the ordering of elements whose placement has not been sufficiently constrained by the grammar.

For a multilingual resource, this allows each language to establish its own balance between the two phases. To show our approach in a nutshell, we present an abstract WO algorithm in Figure 2.

```
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 de-
fault 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
```

Figure 2: Abstract ordering algorithm

The ordering constraints posed by the grammar have the highest priority. Note that this includes the ordering of the textually determined Theme. Then, 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) is restricted by the syntactic structure or (ii) follows SO. The ordering of the CB elements can be (i) specified on the basis of the context, (ii) restricted by the syntactic structure, or (iii) follow SO.

The ordering algorithm as such is not language specific, 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 can be ordered according to information structure. Also, it is desirable (and our algorithm allows it) to specify different systemic orderings for different languages. And, even within a single language, our algorithm allows the specification of different systemic orderings in different grammatical contexts (just by adding a realization statement that (partially) defines the SO during strategic generation).

The algorithm is applicable in platforms other than KPML. In the first place, any grammar can modify its decisions to take information structure into account. In addition, those tactical generators allows multiple sources of ordering constraints, e.g., a combination of grammar-determined choices and defaults, as long as such that the default ordering based on information structure can be applied.

5.2 Algorithm application

The algorithm described above has been implemented and used for generation of Czech and English instructional texts. The Czech grammar resources used in tactical generation have been built up along with Bulgarian and Russian grammar resources as described in (Kruijff et al., 2000), reusing the NIGEL grammar for English. The original NIGEL grammar itself already combines the specification of ordering constraints in the grammar with the application of defaults. If an ordering is underspecified by the grammar, the defaults are applied. The defaults are "static", i.e. specified once and for all. The algorithm we have described replaces these "static" defaults with a "dynamic" construction of 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 SOs for Czech and English specified above (cf. §3.1). 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 3 illustrates this.

Note that the information structure distinction between CB vs. NB elements on the one hand, and the informational status of referents as identifiable vs. non-identifiable on the other hand, are orthogonal. Whereas CB/NB has to do with the speaker's presenting an element as either bearing on the context or context-affecting, identifiability reflects whether the speaker assumes the hearer to pick out the intended referent. These two dimensions are independent, though correlated (cf. the discussion of activation vs. identifiability in (Lambrecht, 1994)). What is encountered most often is the correlation of CB with identifiable and NB with non-identifiable. The correlation of NB with identifiable corresponds is found, e.g., in cases of "reintroducing" an element talked about before, or in cases like *There is a square and a circle. Delete the circle.* –in the second sentence, the same ordering would be used also in German (*Löschen Sie den Kreis*) and in Czech (*Vymažte kruh.*).

```
(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)
```

Generated output:

Pro	otevření	souboru	uživatel	v	menu
for	opening-GEN	file-GEN	user-NOM	in	menu-LOC
vybere	myší	`Open`.			
choose-3SG	mouse-INS	`Open`			

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

Figure 3: Sample input SPL for English and Czech and generated outputs

What is hard to find is the correlation of CB with non-identifiable, but it is the way we would analyze *a dollar bill* in example (9) (Gregory

Ward, p.c.)[4]

(9) (What do you do if you see money laying on the ground?)

Dolarovou	bankovku	bych	zvedla.
Dollar	note	would$_{1sg}$	pick-up$_{1sg}$

Eine	Dollarnote	würde	ich	aufheben.
a	dollarnote	would	I	pick-up

A dollar bill I would pick up.

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

(10)

user	choose	Open
Actor-NB	(Finite-Verb)	Purpose-NB
Uživatel	vybere	pro

menu	mouse	open file
SpaceLoc.-NB	Means-NB	Patient-NB
otevření	souboru	v

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

(11)

user	choose	Open	menu
Actor-CB		SpaceLoc.-CB	(Finite-Verb)
Uživatel	v	menu	vybere

mouse	open file		
Purpose-NB	Means-NB	Patient-NB	
pro	otevření	souboru	myší

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

(12)

user	choose	Open	menu
Purpose-CB		Actor-CB	SpaceLoc.-CB
Pro	otevření	souboru	uživatel

mouse	open file		
Means-CB	(Finite-Verb)	Patient-NB	
v	menu	myší	vybere

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

As mentioned 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 described.

6 Summary and conclusions

We have presented a flexible word ordering algorithm for natural language generation. The novel contribution consists in offering one way of implementing information structure as the major source of constraints on word order variation for languages with pragmatically-determined word order. Apart from that, the special feature of the word order algorithm proposed is that it can also be applied to languages with grammatically-determined word order. We have illustrated the application of the algorithm for Czech and English, Czech being a language in which word order is primarily pragmatically determined and English being a grammatically-determined word order language. We have thus provided evidence that the algorithm can flexibly be applied to 'free' word order languages as well as 'fixed' word order languages.

From a linguistic theoretical point of view, the most important precondition for achieving this has been to take seriously the linguistic observation that in many languages information structure is the driving force for word order variation. For the modeling of information structure for strategic generation, we have drawn upon two well established linguistic frameworks, in both of which the discourse-linguistic and pragmatic constraints on grammatical realization are a focal interest, the Prague School and Systemic Functional Linguistics. From a technical point of view, we have based the implementation on the KPML system, integrating the proposed word order algorithm with existing multilingual grammatical resources and re-using KPML's mechanisms for word order realization as well as its systemic-functionally based notion of Theme. The algorithm is not KPML-specific, though, and could be applied in other frameworks as well, especially if they allow the combination of linearization constraints coming from different sources.

Acknowledgements

The work presented here folows up on our earlier work carried out partially within AGILE (Automatic Generation of Instructions in Languages of Eastern Europe), a project funded by the European Community within the INCO-COPERNICUS programme (grant

[4] Regarding intonation: in English, there are two intonation phrases, the first containing *dollar bill* with a L+H* pitch accent on *dollar*, and the second with a H* pitch accent on *pick up*. In Czech and German it seems that a contrastive pitch accent on *dolarovou bankovku* is optional, and the rest can have neutral intonation with nuclear stress on the last word.

No. PL96114). We would also like to thank the anonymous reviewers of this workshop for valuable comments.

References

John A. Bateman. 1997a. Enabling technology for multilingual natural language generation: The kpml development environment. *Natural Language Engineering*, 3:15 – 55.

John A. Bateman. 1997b. *KPML Development Environment: multilingual linguistic resource development and sentence generation*. Darmstadt, Germany, March. (Release 1.0).

Wallae Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics and point of view. *Subject and Topic*. Charles Li (ed.). New York: Academic Press. p. 25 – 56.

Michael Elhadad and Jacques Robin. 1997. Surge: A comprehensive plug-in syntactic realisation component for text generation. Technical report, Department of Computer Science, Ben Gurion University, Beer Shava, Israel.

Michael Elhadad. 1993. Fuf: The universal unifier user manual 5.2. Technical report, Department of Computer Science, Ben Gurion University, Beer Shava, Israel.

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

Dilek Zeynep Hakkani, Kemal Oflazer, and Ilyas Cicekli. 1996. Tactical generation in a free constituent order language. In *Proceedings of the International Workshop on Natural Language Generation*, Herstmonceux, Sussex, UK.

Michael A. K. Halliday. 1967. Notes on transitivity and theme in English — parts 1 and 2. *Journal of Linguistics*, 3(1 and 2):37–81 and 199–244.

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

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

K. Heidolph, W. Flämig, and W. Motsch. 1981. *Grundzüge einer deutschen Grammatik*. Akademie-Verlag.

Beryl Hoffman. 1994. Generating context-appropriate word orders in turkish. In *Proceedings of the International Workshop on Natural Language Generation*, Kennebunkport, Maine.

Beryl Hoffman. 1995. Integrating “free” word order syntax and information structure. In *Proceedings of the European Chapter of the Association for computational Linguistics (EACL)*, Dublin, Ireland.

Robert T. Kasper. 1989. A flexible interface for linking applications to PENMAN’s sentence generator. In *Proceedings of the DARPA Workshop on Speech and Natural Language*.

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

Geert-Jan M. Kruijff, Elke Teich, John Bateman, Ivana Kruijff-Korbayová, Hana Skoumalová, Serge Sharoff, Lena Sokolova, Tony Hartley, Kamy Staykova and Jiří Hana. 2000. Multilingual generation for three slavic languages. In *Proceedings COLING 2000*.

Ivana Kruijff-Korbayová, John Bateman, and Geert-Jan M. Kruijff. in prep. Generation of contextually appropriate word order. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing*, Lecture Notes. CSLI.

Knud Lambrecht. 1994. *Information Structure and Sentence Form*. Cambridge Studies in Linguistics. Cambridge University Press.

Benoit Lavoie and Owen Rambow. 1997. A fast and portable realizer for text generation. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*, Washington DC.

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

Mark J. Steedman. 1991. Structure and intonation. *Language*, 68:260 – 296.

Mark Steedman. 2000. *The Syntactic Process*. The MIT Press, Cambridge Massachusetts.

Erich Steiner and Wiebke Ramm. 1995. On Theme as a grammatical notion for German. *Functions of Language*, 2(1):57–93.