AGILE

Automatic Generation of Instructions in Languages of Eastern Europe

---

Title ***Modelling Lexical Resources in KPML for Generating Instructions in Slavic Languages***

Authors Geert-Jan M. Kruijff,
John Bateman,
Alla Bémová,
Danail Dochev,
Tony Hartley
Ivana Kruijff-Korbayová
Serge Sharoff,
Hana Skoumalová,
Lena Sokolova,
Kamenka Staykova
Elke Teich,
Jiří Trojánek

Deliverable *LSPEC2*
Status *Final*
Availability *Public*
Date *November 1998*

INCO COPERNICUS PL961104

Abstract:

This document contains the partial results of task 4.2 (lexical and morphological resources for final prototype). In particular, this document contains the deliverable LSPEC2 (specification of a lexical entry). We follow up on the description of the (ideal) lexical entry provided in the deliverable LSPEC1, which sketched the theoretical basis for handling the free word order and the consequences for the lexicon.

The lexical entry specifications we provide here are being used in the construction of the lexicons for the intermediate demonstrator and for the final prototype within task 4.2. They will be described in the deliverables LEXN2-Bu, LEXN2-Cz and LEXN2-Ru. The results of task 4.2 serve as input to deliverable MODL2 (completion of the domain model) and to tasks 7.2 and 7.3 (implementation of generators for the intermediate demonstrator and the final prototype, respectively).

The main problem we address in this report is how to model in KPML lexical resources for Slavic languages, that is to say, how can we indicate for a verb or noun what complements it needs (or may have), and in what form these complements must be realized. Based on data found in the AGILE-corpora for Czech, Russian and Bulgarian, we motivate an approach to modelling lexical resources, and specify prototypically how to implement the model.

More information on AGILE is available on the project web page and from the project coordinators:

URL:	http://www.itri.brighton.ac.uk/projects/agile
email:	agile-coord@itri.bton.ac.uk
telephone:	+44-1273-642900
f	+44-1273-642908

Table of Contents

1. Introduction 1
   1.1 KPML and Natural Language Generation 1
   1.2 Lexical Resources for Slavic Languages: The Issues Involved 2
   1.3 Overview of the Report 2
2. Systemic Functional Linguistics and KPML in More Detail 3
   2.1 Systemic Functional Linguistics 3
   2.2 Systemic Functional Grammar 3
   2.3 The Komet-Penman MultiLingual (KPML) Environment 4
      2.3.1 The Upper Model and the Domain Model 4
      2.3.2 The Grammar as a Network of Systems 6
3. Corpus Investigation 8
   3.1 An Analysis of the Czech Data 8
      Methodology 8
      Results of Corpus Investigation 9
   3.2 An Analysis of the Russian Data 19
   3.3 An Analysis of the Bulgarian Data 25
      3.3.1 Verbs 25
      Мога 31
      Трябва 31
      3.3.2. Substantives 32
      3.3.3. Modifiers 33
      3.3.4. Pronouns 33
      3.3.5. Adverbs, Prepositions and Conjunctions 36
   3.4 Overall Conclusions Based on the Empirical Data 36
4. Modelling Lexical Resources for Slavic Languages in KPML 37
   4.1 Basic Form of Lexical Resources in KPML 37
   4.2 Encoding the Desiderata as Lexical Features 37
      Definition of Additional Lexical Features 38
   4.3 Systemic Concerns 38
   4.4 Prototypical Description of Implementation 39
      Systems dealing with complementations and their morphological form 39
      Systems dealing with possible deletion of obligatory complementations 42

Table of Figures

Figure 1 - Sample domain-related concept 5
Figure 2 - Sample SPL "Draw a polyline" 7
Figure 3 - Abstract form of a lexical entry in KPML 7
Figure 4 Table of Czech verbal and deverbal complementations (directed material processes) 17
Figure 5 - Table of Czech verbal and deverbal complementations (non-directed material processes) 17
Figure 6 - Table of Czech complementations of nouns (corpus data) 18
Figure 7 Table of Russian verbal and nominal complementations (relations) 20
Figure 8 - Table of Russian verbal and nominal complementations (material processes) 22
Figure 9 - Table of Russian verbal and nominal complementations (directed motion processes) 23
Figure 10 Table of Russian verbal and nominal complementations (phase processes) 24
Figure 11 - Table of Bulgarian verbs (material processes) 29
Figure 12 Bulgarian modal verb *мога/ can*, present tense 32
Figure 13 Nominalized forms in the Bulgarian corpus 32
Figure 14 Lexical entries for pronouns 34
Figure 15 Personal pronouns: Bulgarian - English correspondence 35
Figure 16 Personal pronouns: Bulgarian and English correspondences (extended) 35
Figure 17 - Abstract form of a lexical entry in KPML 37
Figure 18 - System for inserting a Goal 40
Figure 19 - Chooser for GOAL-INSERT-CONFLATE 40
Figure 20 - Prototype system for different morphological realizations of a Goal 41
Figure 21 - Revised chooser for GOAL-INSERT-CONFLATE 41
Figure 22 - Location chooser 43

1. Introduction

The overall aim of the AGILE project is to develop a suite of software tools to assist technical writers in the production of manuals for CAD/CAM-software in Bulgarian, Czech, and Russian. The approach taken in AGILE is that of *multilingual generation* from a common semantic representation of the procedural aspects of the tasks involved in using given software tools. The distinctive feature of language generation, as compared to machine translation, is that the meaning of the text is encoded in a formal language rather than a (different) natural language. The platform chosen for developing and implementing the linguistic resources needed for generation is the Komet-Penman Multilingual system, or KPML in short.

This document contains the partial results of task 4.2 (lexical and morphological resources for final prototype). In particular, this document contains the deliverable LSPEC2 (specification of a lexical entry). We follow up on the description of the (ideal) lexical entry provided in the deliverable LSPEC1, which sketched the theoretical basis for handling the free word order and the consequences for the lexicon.

The lexical entry specifications we provide here are being used in the construction of the lexicons for the intermediate demonstrator and for the final prototype within task 4.2. They will be described in the deliverables LEXN2-Bu, LEXN2-Cz and LEXN2-Ru. The results of task 4.2 serve as input deliverable MODL2 (completion of the domain model) and to tasks 7.2 and 7.3 (implementation of generators for the intermediate demonstrator and the final prototype, respectively).

1.1 KPML and Natural Language Generation

KPML is an environment for natural language generation that has its linguistic-theoretical roots in Halliday's Systemic Functional Linguistics (Halliday:1985). Characteristic for KPML is its *semantic* perspective on generation. The idea is that a formal (and unambiguous) specification of the *meaning* of a sentence is given, which is then provided as input to a grammar that is constructed such that it can *realize* a sentence that conveys the intended meaning. This perspective follows naturally from Systemic Functional Linguistics.

Due to its semantic orientation towards generation, and the construction of a grammar as a means to realize meaning, the role lexical resources play in KPML differs from the way lexical resources are usually perceived of in formal grammars like categorial grammar, HPSG, or LFG. When we want to realize a meaning, the grammar creates a representation that delineates the surface form of the sentence by describing the *constraints* the form needs to obey. The grammar construes this representation, or set of constraints, based on the meaning to be realized, without filling in the concrete forms (words) themselves, and thus works relatively independently of a lexicon. When we want to actualize the surface form, by supplying the words that are to make it up, we use words from the lexicon whose lexical description satisfies (or is compatible with) the set of constraints describing the surface form.

1.2 Lexical Resources for Slavic Languages: The Issues Involved

Czech, Bulgarian, and Russian are all languages which are typologically different from English. For one, they all are morphologically rich, particularly when compared to English which is morphologically rather poor. It is this morphological richness which poses one of the major problems for specifying lexical resources and the relation between lexical features and the creation of a constraint-based representation of the surface form.

Namely, we should also be able to specify in a lexical entry for a verb or noun what morphological form(s) its complements are to take. A verb may require, for example, that when the clause it is part of is in active voice and its OBJECT is realized as a nominal group, that nominal group should be in the dative case rather than the accusative case (which would be the default case for realizing an OBJECT as nominal group with a clause in active voice).

Thus, we need to obtain more control not only over lexicogrammatical characteristics pertaining to a verb or noun on its own, but also over characteristics of their complementations. The issues we try to address in this report are therefore:

1. What complementations do the various verbs and nouns take (as found in the corpora), and what morphological forms are these complementations required to have?

2. How can we model these requirements (concerning complementations and their forms) in KPML?

1.3 Overview of the Report

The structure of the report is as follows. We commence by describing the linguistic theory behind KPML, and discuss in detail how this theory materializes in KPML. Subsequently, we present analyses of relevant corpus data for the languages under study in AGILE. The analyses focus on elucidating the actual uses (and occurrences) of verbs and nouns. In particular, the analyses present us with an insight in what kinds of complementations the various verbs and nouns take, and what morphological forms these complementations may be required to have.

Based on the analyses of the data we formulate a number of desiderata that guide our further investigations in how to model lexical resources in KPML for the generation of Slavic languages. We close the report by discussing possibilities for modelling lexical resources, and argue for one particular proposal. This proposal is then worked out in detail, by presenting a prototypical implementation of the necessary semantic and syntactic constructs.

2. Systemic Functional Linguistics and KPML in More Detail

Within the AGILE project, the Komet-Penman MultiLingual system (KPML) provides the platform on which lexico-grammatical resources are developed for the specific purpose of natural language generation. Theoretically, KPML is based on (an interpretation of) Halliday's Systemic Functional Linguistics (Halliday:1985). Before we explain the KPML system, we first briefly discuss Halliday's Systemic Functional Linguistics and its approach to grammar.

2.1 Systemic Functional Linguistics

Systemic Functional Linguistics (SFL) is a linguistic theory developed by Halliday (Halliday:1985), belonging to the continental tradition of functional approaches to the description of natural language. SFL can be characterized by its aim to describe the use[1] of natural language in terms of *functions* and *systems*.

At the highest level of description, SFL considers three broad functions (or rather, metafunctions) of language, being the *ideational* metafunction, the *interpersonal* metafunction, and the *textual* metafunction. The ideational metafunction regards propositional content, whereas the interpersonal metafunction concerns the speakers' roles in (communicative) interaction and their underlying attitudes. The textual metafunction of language has to do with textual organization, in particular the global structure of a text, coherence, and cohesion. SFL holds that all natural languages can be described in terms of these metafunctions.

Each of these metafunctions give rise to a specific kind of meaning - the ideational metafunction construes a model of experience, the interpersonal metafunction enacts social relationships, whereas the textual metafunction creates relevance to context. It is in the grammar that the inputs from the various metafunctions are combined in a process aimed at producing a sentence that reflects these inputs by its structure and choice of words.

Which brings us to the notion of system. As Halliday puts it, a system or *system network* is a theory about language as a resource for realizing meaning. A system by itself represents a choice understood as a set of possible alternatives, be they semantic, lexico-grammatical or phonological. Abstractly speaking, a system includes an entry condition (where the choice is made), the set of possible alternatives, and the realizations (being the structural consequences of each of the alternatives). We can make this picture more concrete by looking at the core component of SFL, namely Systemic Functional Grammar.

2.2 Systemic Functional Grammar

Systemic Functional Grammar (SFG) is an approach to natural language syntax in which grammars are conceived of as networks of systems. The individual systems each reflect a particular aspect of metafunctional meaning and its possible realization. In the process of producing a sentence in which meanings arising from various metafunctions need to be expressed, each system responsible for a (relevant) aspect of meaning imposes specific *constraints* on the form of the sentence. Because systems are networked, the problem of producing a sentence thus becomes one of satisfying a set of constraints. Each constraint

[1] Instead of `use' one could perhaps use the more descriptive though outmoded philosophical term `habit'.

concerns grammatical appearance, ranging from morphological form to word order phenomena.

SFG thus yields grammars which focus on constraints, describing grammatical structure in terms of co-occurrence (co-satisfaction) of those constraints rather than by means of (rewriting) rules[2]. Naturally there are some organizational principles guiding the way in which a grammar is (to be) set up. These organizational principles are *axiality*, *delicacy*, and *rank*.

Axiality expresses the relation between paradigmatic, functionally motivated features and syntagmatic structures realizing them. From a systemic perspective, we find axiality back in the way systems are formulated: technically speaking, a system has input conditions phrased in terms of (grammatical) features, and has as output (grammatical) features, whereby the latter grammatical features may be accompanied by realization statements which connect specific constraints on the realization of the surface form to a particular feature[3]. Delicacy is a principle organizing a grammar in a vertical manner, according to (levels of) specificity. In a network meaning need not be realized immediately, in the sense that there is a single layer of systems. There may be several layers, each successively giving rise to more specific constraints on the eventual realization. The claim of SFG is that languages are (highly) similar in the more abstract layers (grammatical systems of low delicacy) whereas languages tend to differ at the levels of higher delicacy. Finally, rank (and rank scale) expresses a generalized form of a constituency hypothesis. The idea is that a sentence can be divided into clauses, clauses into groups, groups (or phrases) into words, and words finally into morphemes. In other words, we obtain a hierarchy expressing constituency. Parallel to going in a network from a lower delicacy to a higher delicacy we can usually observe a move from a higher rank to a lower rank: the higher the delicacy, the closer we get to actual realization. Conform to the claim that languages tend to differ at levels of higher delicacy, we see that different languages may have different preferences concerning the rank at which a certain phenomenon is expressed.

2.3 The Komet-Penman MultiLingual (KPML) Environment

The KPML system, short for *Komet-Penman MultiLingual* system, is a platform for implementing grammatical resources based on the Penman system for generating English. As pointed out above, the linguistic theory underlying KPML is Halliday's Systemic Functional Linguistics.

2.3.1 The Upper Model and the Domain Model

In KPML, the ideational meaning (propositional content) is an instance of what is called the *Upper Model* (or UM for short). The UM provides a general organization of knowledge. Knowledge pertaining to a particular working domain is modelled as a specialization of the UM, employing basic concepts defined there. In AGILE, the working domain is computer-aided design and manufacturing (CAD/CAM), and the domain model (DM) specifies the concepts that together provide the background knowledge needed for generating explanations in this domain.

For example, in the AGILE DM, we can find the following domain-related concepts (cf.

---

[2] A well-known approach to grammar which is also constraint-based is HPSG (Pollard & Sag:1993).

[3] See also the section on KPML and its notion of grammar, below.

WP2-1,p.5):

```
(DEFINE-CONCEPT dispositive-material-action (directed-action)
       ((ACTEE :TYPE OBJECT)))
(DEFINE-CONCEPT enter (dispositive-material-action)
       ((LOCATION :TYPE SPATIAL :OPTIONAL T)))
```

Figure 1 - Sample domain-related concept

The concept *dispositive-material-action* is a subtype of the concept *directed-action*, and it has as slot an *actee* which is of type *object*. Because we consider slots as obligatory unless defined optional, the actee slot is defined as obligatory for this concept. The concept *enter* is a subtype of *dispositive-material-action*. *enter* itself defines a *location* slot, of type *spatial*, which is optional[4]. Due to inheritance, the *enter* concepts has not only a *location* slot, though, but also an *actee* slot - an inquiry

```
(get-concept-slot-descriptions 'enter)
```

would evaluate to

```
((actor:obligatory) (location spatial :optional)(actee
                  object :obligatory))
```

the actor is inherited from *material-process* via *directed-action*.

As a matter of fact, a closer look at the AGILE DM reveals that the concepts employed are organized in a type-subtype hierarchy. An example is the *enter* concept above. Other concepts defined are *print, quit-tool*, etc. For each of these concepts, slots are defined that need to (or may) be filled and which correspond to (ideational) functions.

As Bateman *et al* describe in (Bateman et al:1990), there is a close relation between the upper model and the grammar realizing propositional content, in the following sense. When we want to generate a sentence, we create a specification of the content the sentence is to express, based on the UM and the DM. The specification is in the form of a so-called SPL expression (SPL) (which also states interpersonal information (notably, speech acts) and textual information, besides propositional content).

The grammar is constructed such that it is capable of realizing forms for individual concepts and relations in the UM: it consists of a paradigmatic description of linguistic phenomena (corresponding to the UM) and a syntagmatic description of their realisation.

The grammar's paradigmatic description of linguistic phenomena follows SFG's rank scale. The rank scale gives the basic paradigmatic grammatical classes for which particular sets of systems and their features hold, and it defines the basic constituency organization of syntagmatic structure (op.cit. WP6-1, p.2). The claim of KPML, and SFL, is that the paradigmatic description is more likely to be shared across languages than the syntagmatic description.

[4] ":OPTIONAL T" means that the optionality-attribute is set to true.

2.3.2 The Grammar as a Network of Systems

Let us have a more precise look at how the grammar component of KPML works. To begin with, we should clarify what the idea behind grammar and grammatical structure is, in KPML as well as in systemic-functional grammar in general. The gist of the matter is that, depending on the meaning to be actualized, the grammar makes a choice for specific *grammatical features* that characterize the surface form as a whole. These grammatical features by themselves are not directly concerned with constituency structure, though. They are related to the surface form by means of *realization statements* which are associated to grammatical features.

Realization statements are defined in terms of *grammatical functions* and operations on these functions. A grammatical function describes the function which a particular constituenct is to perform - conceiving of grammatical structure as a configuration of grammatical functions, it is thus in this way that we actually build a set of constraints delineating the realizable surface form. The operations on grammatical functions, which are also called *realization operators*, can be grouped into three functional categories:

1.Operators defining particular grammatical constituents - for example, by *insertion* of a grammatical function *Subject*, by *conflation* of a grammatical function *Actor* with the function *Subject*, or by *expansion* of a constituent as belonging to a higher function.

2. Operators imposing linear ordering constraints on constituents - for example, *Partition, Order, OrderAtFront* and *OrderAtEnd* - are realization operators imposing ordering constraints.

3. Operators that associate features with functions, and which are as such concerned with how constituents are to be realized rather than with the specification of constituents as playing particular functions.

Recapitulating, when we make a traversal through the network of systems, particular grammatical *features* get selected as means to reveal particular aspects of meaning. These features are, in turn, associated to *realization statements* which constrain the surface form the grammar is producing. Realization statements delineate configurations of grammatical *functions* and the *constituents* that realize them.

Technically speaking, in KPML each system in the network has associated to it a so-called *chooser* which consults with the UM and the current SPL to decide which grammatical features to select to realize aspects of the meaning specified in the SPL. A chooser is a decision tree in which each node is an *inquirer*. It is the inquirers that in effect interpret semantic information. Based on the outcome of the chooser, the system selects a grammatical feature which may have associated realization statements. Thus, a chooser mediates between semantic and grammatical information.

Going down the network, following the rank scale, constraints may get imposed on the realization that specify the actual form of the sentence in more and more detail. The actual path through the network naturally depends on the choices made earlier, or higher up, in the network.

An interesting aspect of this kind of grammar is that it is *semantically* oriented, and is relatively independent from lexicogrammatical information specified for individual words. In fact, if an SPL does not specify by what words a particular meaning needs to be realized, the structure resulting from the traversals through the network specifies a syntactic structure including all the morphological constraints on the words, without the words themselves. To

fill in words an additional *lexify* step is needed, and subsequently a call to a morphological component to “really“ realize the sentence (an *inflectify* step).

Naturally we can already specify in an SPL what words are to be used, so that at the end of the traversal only the morphological component needs to be called. Take for example the next SPL:

```
(S / dispositive-material-action :LEX draw
     :SPEECHACT command
     :ACTEE (D / OBJECT :LEX polyline
                :IDENTIFIABILITY-Q non-identifiable
 ))
```

Figure 2 - Sample SPL "Draw a polyline"

The SPL defines a command involving a non-identifiable (that is, arbitrary) ``polyline" that is to be ``drawn" - which would result in the English sentence ``Draw a polyline". Regarding non-identifiability of the object, IDENTIFIABILITY-Q corresponds to an inquirer in the chooser deciding about the realization of the object. Here, we explicitly prescribe the answer to that query.

The :LEX labels specify which lexical items (words) should be used in the realization of the sentence. Abstractly, a lexical entry for a word is of the following form:

```
(lexical-item
      :NAME name-of-the-item
      :SPELLING ``spelling of the item''
      :SAMPLE-SENTENCE ``a sample sentence using the item''
      :FEATURES (lexical-feature-1 .... lexical-feature-n)
      :PROPERTIES (property-1 ... property-m)
      :EDITOR ``name of the editor''

)
```

Figure 3 - Abstract form of a lexical entry in KPML

The :NAME in the lexical entry specifies a label, by which the entry can be referred to by a :LEX statement in an SPL. The :FEATURES of the item describe the lexicogrammatical character of the word in terms of *lexical features*. In a *lexify* step, these lexical features act as constraints which need to be satisfiable in the context of the set of constraints formed by the grammar as a delineating of the surface form.

3. Corpus Investigation

In this section, we present analyses of the AGILE corpora relevant to the current stage of the project and especially for the task of specifying the format of a lexical entry. The main focus was to find out what requirements various verbs and nouns have on the realization of their complementations.

Results of analyses are presented for the three Slavic languages under examination in AGILE, being Czech, Russian, and Bulgarian. Although we tried to unify the methodology employed for the present investigation by the individual groups as much as possible, there are still some differences concerning the exact approaches taken. Each group describes the approach used in the respective section.

3.1 An Analysis of the Czech Data

Methodology

For the analysis of the Czech corpus, we employ an approach to describing verbal and nominal complementations based on the notions of dependency relations and valency frames. The approach has been developed in Prague over the last two decades in relation to the framework of Functional Generative Description (Sgall et al 1986). The main idea is that we describe how a verb or noun can act as a head by specifying the dependency relations that can modify it. Together these dependency relations make up a valency frame. The notion of valency frame is comparable to that of the subcat list in HPSG or the θ-frame in Government & Binding theory.

The relation between heads and valency frames in terms of dependency relations on the one hand, and the concepts and slots of KPML on the other hand, can be briefly described as follows. We can regard a valency frame and its dependency relations as a syntactic description of the grammatical construction realizing the concept and its slots. Even though there is not an isomorphic mapping between the slots and the dependency relations in all cases, it can be assumed that there is a systematic relation. For instance, the slots Actor and Actee are isomorphic to the dependency types Actor and Patient, but the slot Location can correspond to the dependency type of either Location or Origin. Another point of comparison between concept slots and dependencies is that the coverage of the latter is broader, since it also includes relations which are not captured as slots of a concept, e.g. Manner (how something is accomplished), Effect (the result of some action), etc.

The theoretical aspects of the Praguian approach have been described in the works on verb valency (Panevová 1974-1975, 1977, 1978; Hajičová 1979, 1983). The main points and terminology can be summarized as follows:

- Among the **complementations** of verbs we distinguish between **inner participants** and **free modifiers**.
- The **valency frame** of a given verb includes some of the inner participants and those free modifiers which fulfil the criteria of being semantically obligatory.
- **Inner participants** can be the following: Actor (Act), Patient (Pat), Addressee (Addr), Origin (Orig), Effect (Eff).[5] Inner participants can be obligatory of facultative

[5] The particular terminology is of course not as important as the functions that the complementations fulfil

(optional).

- **Free modifiers**, e.g. modifiers expressing time, location, manner-means, reason (why), goal (purpose) etc., can occur with any verb. However, with some verbs they are semantically obligatory. In these cases they belong to the valency frame of the verb, although they are often deletable in the text. Examples are: the modifier expressing location-direction which means "where-to" with the verb *come*, the modifier expressing location-source which means "where-from" with the verb *leave*, or the modifier expressing location-place which means "where" with the verb *appear.*

With respect to the usual realization of the inner participants, we can say the following for Czech, and also for other languages with inflection we are familiar with: in a sentence in active voice, the Actor is usually realized by a nominal group in the nominative case (grammatical Subject), for the Patient it is the accusative case (grammatical direct Object), and for the Addressee it is the dative case (grammatical indirect Object). For the Origin and Effect there is no single prototypical form. In Czech, for instance, the Origin is commonly realized by a prepositional group consisting of *z (from)* or *od (from)* and a nominal group in the genitive case, for the Effect it is *do (to, into)* and genitive, *na (on, onto)* and accusative or *v (in, into)* and accusative. There also exist exceptions, and therefore it is not easy to prescribe rules for the realization of free modifiers.

Results of Corpus Investigation

We analyzed the Czech AGILE corpus (see the deliverable CORP-Cz) following the principles overviewed above. Our observations on verbs can be summarized as follows:

- **Actor (Act)**: present in the valency frames corresponding to all the verbs in the corpus; however, it is scarcely realized in the texts, mainly due to the fact that most of the verb occurrences are in the imperative mood in which the Actor is not realized. Also in the declarative mood, the Subject is often not present on the surface if the Actor and the Subject collide and refer to the user in a direct way, i.e., by a pronoun in first or second person if it were in English (Czech is a Subject pro-drop language). When realized, the Actor collides with Subject in active voice, and it is in the nominative case.
- **Patient (Pat)**: all but 3 verbs have a Patient in their valency frames; the exceptions are *objevit se (appear)*, *pokračovat (continue)* and *začít (begin* or *start*, in one reading*)*; the Patient is usually realized by a nominal group in the accusative case in active voice; there are the following exceptions: *pohybovat (move)* takes a Patient in the instrumental case, *začít (begin* or *start*) in accussative or as an infinitival clause (e.g., *začít fungovat (start functioning)*), and the Patient of *navázat (continue)* is a prepositional group with the preposition *na (on)* and a nominal group in the accusative case: *navázat na čáru (continue a line);* There are two cases of ellipsis in the corpus, both within the same sentence, where the Patient which is in the valency frame of a verb is not realized: *Nyní pojmenujte a uložte.* (*Now name and save*. -- segment No. 131 and 132 in the Czech corpus, part of „Vytvoření stylu multičáry" (Creating multiline style, p. 47))
- **Origin (Orig)**: included in the valency frames of the verbs *vybrat (select)*, *zvolit (choose)*, *vytvořit (create)*; it is realized by a prepositional group with the preposition

in the meaning of a sentence.

z (from) and is facultative, and therefore deletable in the text

- **Effect (Eff)**: included in the valency frames of the verbs *spojit (connect)*, *měnit (change*, irreflexive, imperfective) and *změnit se* (*change*, reflexive, perfective); with the verb *spojit*, the Effect is realized by a prepositional group with the preposition *do (into)* and a nominal group in the genitive case: *spojit do páru (connect into a pair)*, with the verbs *měnit* and *změnit se* Effect is realized by a prepositional group with the preposition *na (into* in this case*)* and a nominal group in the accusative case: *(z)měnit na něco (change into something)*; both *měnit* and *změnit* have a Patient in the valency frame, but *změnit se* is a reflexive form, so there is the Patient is realized by the reflexive particle *se* in the sentence
- **Manner/Means (Man)**: it is not an inner participant, and it does not belong to the valency frames of any of the verbs in the corpus, however, it appears with the following ones: *editovat (edit), nastavit (set* or *adjust), otevřít (open), přepínat (switch), vymazávat (delete)* and *spustit (start);* when it is realized in the corpus, it is by a nominal group in the instrumental case, sometimes in the form *pomocí něčeho (using something)*. There is also the form "*jedním z následujících způsobů*" (*using one of the following methods*) in instrumental case which appears in almost every instruction set
- **Location (Loc)**: it is not an inner participant, but it does belong to the valency frames of the following verbs in the corpus: *objevit se (appear)* takes a Location where (Place, Plc), *uložit (save)*, *umístit (place* or *position)* and *zapsat (save* or *write)* take a Location where-to (Direction-to, Dir-to); although it belongs to the valency frame, Location is deletable with all these verbs in Czech, so it does not have to be realized in the sentence

Our corpus also contains many occurrences of deverbatives, i.e. nominalizations. Such "nouns" inherit the valency frame of the verb from which they are derived. As for their realizations, it holds in general that the Actor, if it should be realized, would be realized by a possessive construction. The Patient, if it is realized by a nominal group, takes the genitive case (instead of the usual accusative with the verb) – we have not encountered any exceptions to this general rule in our corpus. The other cases remain the same as with the corresponding verb.

Also ordinary nouns may have valency frames. In our corpus we have encountered nouns with a Patient in the valency frame.

The reason why we employed the approach using valency frames in our analysis is two-fold. On the one hand, we would like to follow up on the Praguian tradition of describing Czech syntax, and in order to do this, we need to work with the notions we are used to work with, and relate them to the SFL framework and to the KPML notions in particular. On the other hand, we believe that the corpus analysis we carried out using valency frames rather than KPML notions of various types of processes and their "roles" is more general. It was our intention not to restrict ourselves to the KPML notions and classifications already present in the current grammar, because we believe that in this way we have a better chance to determine what changes we need to make in the grammar systems in order to adapt them for generating Czech. An example in this context is the distinctions we draw between the dependency types of Origin versus Location or Direction, and between Effect versus Direction. When looking only at their realizations, one probably would not make these distinctions: both Origin and Location can be expressed by a prepositional group with the

preposition "*v" (in)* and the locative case, Origin and Direction can be expressed by a prepositional group with the preposition "*z" (from)* and the genitive case, Effect and Direction can both be expressed by a prepositional group with the preposition "*do" (to, into)* and the genitive case. However, while for Origin and Effect, which we consider to be among the inner participants in the valency frame of a verb, the mentioned forms are the only ones available, Location and Direction are ordinary free modifiers (called circumstances in Halliday's SFG) that can be expressed in a number of ways as adverbial modifications, using prepositional groups as well as adverbs. The latter are not available as means expressing Origin (e.g., "*vybrat v menu" (choose in a menu)*) or Effect (e.g. "*spojit do páru" (connect (in)to a pair)*).

We believe that it is a good strategy for the development of the Czech grammar to distinguish among different types of dependencies and project these distinctions into the KPML system network. In this way we are able to impose grammatically appropriate restrictions on the realization of verb and noun complementations in various contexts (this is our task in Work Packages 6 and 7, concerned with linguistics specifications and their implementation).

The observations we made in the analysis of the AGILE corpus are summarized in the tables that follow. Some general remarks concerning the contents of the tables will be made here. Remarks pertaining only to each of the individual tables will be made below.

First of all, a difference between English and Czech (as well as other Slavic languages) is that functions that are purely grammatical in English are realized by lexicogrammatical categories in Czech. This concerns such grammatical phenomena as aspect and deverbalization (nominalization), which are lexicalized in Czech, i.e. there are separate word-forms for different aspect of the same verb base-form. One can deverbalize each of them, thus obtaining two different nominalizations. For instance, the verb occurring most often in the AGILE corpus is "*spustit*" (start), which is of perfective aspect; the corresponding imperfective form is "*spouštět*". The corresponding two nominalizations are "*spuštění*" and "*spouštění*", respectively.

There are various means of creating perfective verb forms out of imperfective ones and the other way round in Czech, using infixing or prefixing. For example, the verb "*měnit*" (*change*) is imperfective. There is a number of perfective verbs which can be considered derived from it by prefixing: "*změnit*", „*přeměnit*", „*vyměnit*", „*zaměnit*". From the latter three, we can derive imperfective forms again, by infixing: "*přeměňovat*", „*vyměňovat*", „*zaměňovat*".[6]

Therefore, it would be quite a hopeless task to try to enumerate all the possible alternatives which could be derived to complement a verb form that actualy appears in the corpus. Therefore, we do not do this, and only mention the verb base-forms of the verbs actually encountered in the corpus. In case we also encountered the corresponding nominalization, we include it in the table, otherwise not. There are some cases where we only encountered the nominalization in the corpus; in these cases, we also include the verb base-form from which the nominalization is derived. As mentioned above, the nominalizations „inherit" the valency frames from the verbs, so we present the verbs and their nominalizations in one line in the tables below.

[6] Except for the differences in aspect, měnit, změnit, přeměnit, přeměňovat appear synonymic and have the same valency frame.

Each table contains several columns. First, there is the word-base and the English gloss for it. Then we include the corresponding AGILE domain model concept and its slots, if the concept has already been defined.[7]

Furthemore, we present the valency frame according to the FGD theory. Obligatory complementations (inner participants or free modifiers) are stated without brackets, the optional ones (inner participants) are enclosed in brackets; free modifiers other than the obligatory ones can always be optionally added, so they are not explicitly mentioned in the valency frames. Note that an obligatory free modifier can usually be deleted in the sentence. We use the dependency types in terms of valency frames also in the column which captures the corpus occurrences. This column reflects the actual realizations of complementations as found in the corpus, along with the main characteristic of the occurring morphological form. If a complement is realized, it is stated in the occurrences patterns; the cases where it is not realized, are separate cases. This is why we for instance distinguish an accurrence of „Pat-acc“ and „Pat-acc Loc-v+loc“.We use abbreviations for cases (nominative, genitive, dative, accusative, locative, instrumental); when the realization is restricted to a particlar form of a prepositional group or only a particular form appears in the corpus, we state the preposition and the case required by it; when the form is not restricted, we just say „adv“ for adverbial, which can be a prepositional group or an adverb.

We state all realization information that we consider important for the Czech grammar development in AGILE. This concerns the case, active vs. passive distinction and specific realization forms, e.g. “pomocí+acc”, etc.

Note that we are trying to capture our observations here; we may observe phenomena that are already implemented in the KPML grammar, but we may also observe ones that are not. This is the reason why we do not use realization statements to describe the corpus occurrences. In fact, we will be formulating realization statements and other pieces of the KPML implementation within Work Packages 6 and 7 (specification and implementation of linguistic resources) on the basis of the observations made in the corpus analysis presented here.

The last column in each table contains the numbers of segments in the corpus where the given word has been encountered (it is not necessarily an exhaustive listing).

We distribute the results of the corpus investigation of verbs and their nominalizations over a number of tables, depending on the process type from the KPML upper model to which we expect to relate the given word-base. In addition, we include a table containing nouns which also appear with particular types of complementations in the AGILE corpus, and therefore can be seen as having valency frames.

The table in Figure 4 contains a survey of our corpus observations concerning verbs and nominalizations expressing directed material processes. The table in Figure 5 contains a survey of our corpus observations concerning verbs and nominalizations expressing non-directed material processes. Of the verbs expressing relations, we have only encountered occurences of the verbs „být“ (be) and “mít” (have).

Word-base	English	DM concept	DM	Valency	Corpus	Corpus

[7] Deliverable MODL2 will complete the definitions of concepts needed for the intermediate and final prototypes.

	Gloss		Concept Slots	frame	Occurrences	References
definovat	define	located-action ENTER	Actor Actee Location	Act Pat	Pat-acc	47,87
dokončit dokončení	finish	simple-action END	Actor Actee	Pat	 Pat-gen	 281
editovat	edit	data-action EDIT	Actor Actee	Act Pat	Pat-acc Man-“pomocí”+gen	70,207,278
měnit, přeměnit, změnit přeměnění	change	simple-action CHANGE-COMPONENT	Actor Actee	Act Pat (Eff)	Pat-acc Pat-acc Time Pat-acc Eff-“na”+acc Pat-gen Eff-“na”+acc	72,74,282 72
nastavit nastavení	set	simple-action SPECIFY COMPONENT	Actor Actee	Act Pat	Pat-acc Pat-acc Eff[8] Pat-gen	26,256 261 138,258
navazovat	continue			Act Pat	Pat-“na”+acc	45
objevit se	appear			Act Loc	Loc-where	104,126
opakovat	repeat			Act Pat	Pat-acc	101
opsat opsání	circumscribe	simple-action DRAW	Actor Actee	Act Pat	 Pat-dat	 204
opustit opuštění	leave	simple-action QUIT TOOL	Actor Actee	Act Pat	 Pat-gen	 99
otevřít	open	data-action OPEN or simple-screen-action OPEN-SCREEN-OBJECT	Actor Actee Actor Actee	Act Pat	Pat-acc Man-jeden	79
pohybovat	move			Act Pat	Pat-instr Pat-instr Loc-where Pat-instr Loc-Dir-to	226 268 270

[8] The verb group is “nastavit proměnnou na hodnotu” (set the variable to a value).

Word-base	English Gloss	DM concept	DM Concept Slots	Valency frame	Corpus Occurrences	Corpus References
pojmenovat	name	simple-action SPECIFY COMPONENT	Actor Actee	Act Pat	Pat-acc	131
posunout posunutí	shift, move			Pat	 Pat-gen	 86
použít	use			Act Pat	Pat-acc Pat-acc Goal-"pro"+acc	245 71
přehlížet	overlook			Act Pat	Pat-acc	250
překřížit	cross			Act Pat	Pat-acc	112
přepínat přepnutí	switch			Act Pat (Orig) (Eff)	Pat-acc Man-"pomocí"+gen Eff-"do"+gen Orig-"z"+gen	247 59 249
přidat přidání	add			Act Pat Dir	 Pat-gen Dir-"k"+dat Pat-gen	 85,90 137
přijímat	accept			Act Pat	Pat-acc	223
rozkládat rozkládání	decompose			Act Pat	 Pat-gen	 73
spojit spojení	connect			Act Pat (Eff)	Pat-acc Pat-acc Eff Pat-gen,"s"+instr[9]	114,118,121 113 14
spustit	start	simple-action START TOOL	Actor Actee	Act Pat	Pat-acc Pat-acc Man-instr Pat-acc Man-"pomocí"+gen Pat-acc Man-"jeden"+gen[10]	212 2,28,52 195 148

[9] The nominal group "*spojení něčeho s něčím*" (connecting something with something) can be seen as a noun modified by a complex Patient consisting of two coordinated nominal groups. The coordination takes the form of an "s"+instr (with).

[10] There are three different ways of realizing Manner (Means) in Czech mentioned here, in order to make these alternatives explicit for the later phase of grammar development.

Word-base	English Gloss	DM concept	DM Concept Slots	Valency frame	Corpus Occurrences	Corpus References
spustit	lower			Act Pat	Pat-acc Dir-adv	219
stisknout stisknutí	enter, press	simple-action PRESS	Actor Actee	Act Pat	Pat-acc[11] Pat-gen	11,37,218 18
ukončit ukončení	end, terminate	simple-action QUIT-TOOL	Actor Actee	Act Pat	 Pat-gen	 2,15,28,68
uložit uložení	save	data-action SAVE	Actor Actee	Act Pat Dir	Pat-acc Pat-gen Pat-gen Dir-"do"+gen	132,277 98 140
umístit	place			Act Pat Dir[12]	Pat-acc Man-"podél"+gen	75
určit určení	specify, determine			Act Pat	Pat-acc Pat-gen	7,106,166,187 186,201
uzavřít uzavření	close	simple-action QUIT TOOL	Actor Actee	Act Pat	Pat-acc Pat-gen	142 175
vepsat vepsání	inscribe			Act Pat Dir	 Pat-elided Dir-"do"+gen	 189
vrátit vrácení	return			Act Pat Dir	 Pat-gen Dir-adv	 15,275
vybrat výběr	select, choose	select-action CHOOSE	Actor Actee Options	Act Pat (Orig)	Pat-acc Pat-acc Orig-"z"+gen Pat-acc Goal-"pro"+acc Pat-acc Goal-"k"+dat Pat-gen	84,89,107,124 4,81,152,214 84,89 137 96

[11] The Patient of stisknout is either a name/label of a button or a key, or it is a specific description, e.g. tlačítko ENTER.

[12] The verb group is "*křivka je umístěna podél středu"* (line is placed along the centre), in which the prepositional group can be considered to express Manner, i.e. how the line is placed.

Word-base	English Gloss	DM concept	DM Concept Slots	Valency frame	Corpus Occurrences	Corpus References
vymazat vymazávat vymazání	delete, erase			Act Pat	Pat-acc Pat-acc Man-"pomocí"+gen Pat-gen	271 279 263
vyplnit vyplnění	fill in			Act Pat	Pat-gen	128
vypsat vypsání	write-out, display	data-action PRINT	Actor Actee	Act Pat	Pat-gen	156
vytvořit vytvoření	create			Act Pat (Orig)	Pat-acc Pat-gen	25,41 78,100
začít	begin, start	simple-action START-TOOL	Actor Actee	Act Pat	Pat-acc[13] Loc-where Man-instr	17
zadat	enter, specify	simple-action SPECIFY COMPONENT	Actor Actee	Act Pat	Pat-acc Pat-acc Goal-"pro"+acc[14] Pat-acc Condition-when[15] Pat-acc Condition-if[16]	64,66 12,16 153 265
zapsat	write, save	data-action SAVE	Actor Actee	Act Pat Dir	Pat-acc Dir-"do"+gen	231
zarovnat zarovnání	align			Act Pat	Pat-gen	160
zaznamenat	record	data-action SAVE	Actor Actee	Act Pat	Pat-acc	284
zmáčknout	press	simple-action PRESS	Actor Actee	Act Pat	Pat-acc	158,171

[13] The Patient of začít can also be realized by an infinitival complement, although no such occurrence has been encountered in the AGILE corpus.

[14] The Patient of *zadat* is either a specific value, *e.g. u, z, ?*, or it is a "generic" description, e.g. *vzdálenost (distance), úhel (angle), délka (length).* The purpose is never realized when the Patient is a generic description, When the Patient expresses a specific value the Purpose is realized in almost all cases.

[15] The expression is a prepositional group "*při dotazu na styl" (at the query for style).*

[16] The condition is expressed by an if-clause.

Word-base	English Gloss	DM concept	DM Concept Slots	Valency frame	Corpus Occurrences	Corpus References
zobrazit zobrazení	display	simple-action	Actor Actee	Pat Loc	 Pat-gen Pat-gen Loc-"v"+loc	 109,125
zvolit	choose	select-action CHOOSE	Actor Actee Options	Act Pat (Orig)	Pat-acc Pat-acc Orig-"z"+gen[17]	91,94,133 93,96,128

Figure 4 Table of Czech verbal and deverbal complementations (directed material processes)

Word-base	English Gloss	DM concept	DM Concept Slots	Valency frame	Corpus Occurrences	Corpus References
kreslit kreslení	draw	simple-action DRAW	Actor Actee	Act Pat	Pat-acc Pat-gen	1,27,59,63 22,232
nakreslit nakreslení	draw-perfective	simple-action DRAW	Actor Actee	Act Pat	Pat-acc Pat-gen	44,88 50,147,179
obnovit obnovení	renew, resume			Act Pat	Pat-acc Pat-gen	229 238
pokračovat	continue			Act Pat Act[18]	Pat-"v"+loc	235
začít	begin			Act	Time	240
zdvihnout se	lift			Pat	Act-nom Dir-adv Act-nom Man	267,225

Figure 5 - Table of Czech verbal and deverbal complementations (non-directed material processes)

[17] This is really Origin, not Location-from, due to the restricted repertoir of realizations, e.g. we cannot say "zvolit od" or "zvolit zpoza", which we could, if it were a Location.

[18] The only occurrence of Actor-only valency frame encountered in the AGILE corpus.

In addition, also nouns can be considered to have valency frames. We have compiled a table of nouns and their complementations which exhibit interesting behaviour in this respect in Figure 6. In particular, we do not include nominal complementations realized by a nominal group in the genitive case, because it appears that every noun can have such post-modifier (corresponding to the English "of" construction).

Wordbasis	**English Gloss**	**Theoretical valency frame**	**Corpus Occurrences**	**Corpus References**
dotaz	question	Pat	Pat-"na"+acc	19,153,158
návrat	return	Dir-to	Dir-"do"+gen	63,130,273
pohyb	move	Pat	Pat-Instr	220
vzdálenost	distance	Pat	Pat-gen,"od"+gen[19]	64 146
záznam	record	Pat Manner Pat	Pat-gen Man-"od"+gen Pat.-gen	211 244

Figure 6 - Table of Czech complementations of nouns (corpus data)

[19] The nominal group has the form "vzdálenost něčeho od něčeho" (distance of something from something), which can be seen as a complex Patient consisting of two coordinated nominal groups.

3.2 An Analysis of the Russian Data

The analysis is based on the notions employed in SFL; in particular we use the concepts in the AGILE Domain and Upper Models and functions and features used in the current version of the Russian grammar in KPML. In our analysis we do not distinguish Participants and Circumstances in the valency frame of verbs. They depend on semantics of the situation and are specified in the process of the translation from A-box to SPL.

The tables in Figure 7 through Figure 10 contain a detailed overview of the verbs found in the AGILE corpus. The former shows relations, the latter shows material processes. Some comments concerning to the content of its columns follow.

A significant difference between English and Russian is that functions that are purely grammatical in English are realized by lexicogrammatical categories in Russian. So such grammatical phenomena as aspect and nominalization are lexicalized in Russian, i.e. the opposite-aspect verb or the substantive for a verb should be realized by another word, often having some additional connotations in its meaning and sometimes not existing at all (thus complicating text planning tasks).

These three forms, namely perfect and imperfect verbs forming the aspectual pair and their nominalization(s), are shown in the first column of the tables in Figure 7 through Figure 10, namely the column **Wordbasis.** Elements which are not present in our corpus, but available in the system of language, are shown in brackets. The second column, **English Gloss** in the tables contains an English equivalent(s) for this lexical nest as they occur in the corpus (thus restricting all possible meanings). The columns **Domain Model Concept** and **Domain Model Concept Slots**, when included, contain the corresponding information from the AGILE domain model (if the corresponding concept already exists in the current version of the T-box).The column **Upper Model Process Type** contains KPML Upper model concepts (we added three new concepts to the existing Upper Model). The column **Corpus Occurrences** contains the set of complements of the verb as they occur in the corpus. They are expressed in terms of UM roles with added realization constraints imposed by the grammar. Optional complements are enclosed in brackets. The last column contains the numbers of segments in the corpus in which the given verb occurs (it is not necessarily an exhaustive listing).

Comments specific to Figure 7: These verbs are semantically relations. They are universal with respect to a problem domain. Probably in the GUM we need a concept of forming as a subclass of identity (a source by which the identified is formed). As for the Russian UM: ways for expression of identity differ in their government pattern.

Comments specific to Figure 8 and Figure 9: We propose some additions to the Upper Model. The first is the notion of lexical function verbs, for example, *позволять* (allow)+*inf* corresponding to an English finite form in cases of stylistic problems in using Subject as Actor in order to lower it to Means for an action:

PLINE draws poliline segments...
Команда PLINE позволяет рисовать сегменты полилиний...

One more lexical function verb is *хотеть* in:

Нажимайте клавишу r каждый раз, когда хотите записать...
Enter r at any time to record...

It is introduced to express the purpose in the finite dependent clause replacing the infinite clause in English. Our corpus lacks significant number of such verbs, though they are quite common in Russian (Mel'cuk's Oper function and its offsprings).

Wordbasis	English Gloss	UM process type	Corpus Occurrences	Corpus References
(-) являться (-)	be	identity	Domain Nom Range Instr	40
образовать[20] (образовывать) (-)	form	identity	Domain Nom Range Acc	22
(-) состоять (-)	combination	part-whole	Domain Nom Range «из» + Gen	42

Figure 7 Table of Russian verbal and nominal complementations (relations)

Wordbasis	English Gloss	DM concept	DM concept slots	UM process type	Corpus Occurrences	Corpus References
ввести (вводить) (ввод)	enter	Located-action ENTER	Actor Actee Location	dispositive - material-action	Actor Actee (Static-spatial) «в»+ loc	69, 115, 153
(задать) задавать (задание)	set			dispositive - material-action	Actor Actee	25
закрыть (закрывать) (закрытие)	exit close			dispositive - material-action	Actor Actee	80, 107
замкнуть (замыкать) (замыкание)	close			dispositive - material-action	Actor Actee	38

[20] The passive form is available in which it can be synonymous to "состоять".

Wordbasis	English Gloss	DM concept	DM concept slots	UM process type	Corpus Occurrences	Corpus References
записать записывать (запись)	record	Data-action	Actor Actee	dispositive - material-action	Actor Actee (Destination) «в/на» + acc	175, 191
сохранить сохранять сохранение	save	Data-action SAVE	Actor Actee	dispositive - material-action	Actor Actee (Destination) «в/на» + acc	79, 98
запустить запускать запуск	start	Simple-action START	Actor Actee	dispositive - material-action	Actor Actee (Means) + Instr	2, 14, 18, 27
нажать нажимать нажатие	press enter choose select click	Simple-action PRESS	Actor Actee	dispositive - material-action	Actor Actee	11, 17, 19
описать (описывать) (-)	circum-scribe			dispositive - material-action	Actor Actee	157, 170
определить (определять) (определение)	specify	Simple-action SPECIFY	Actor Actee	dispositive - material-action	Actor Actee	84
опустить (опускать) (-)	put down			dispositive - material-action	Actor Actee	184
отменить (отменять) (отмена)	undo			dispositive - material-action	Actor Actee	16, 134
открыть (открывать) (открытие)	open	Data-action OPEN	Actor Actee	dispositive - material-action	Actor Actee	61
поднять (поднимать) (поднятие)	lift up			dispositive - material-action	Actor Actee	186

Wordbasis	English Gloss	DM concept	DM concept slots	UM process type	Corpus Occurrences	Corpus References
показать (показывать) (показ?)	display show list	Simple-action DISPLAY	Actor Actee	dispositive-material-action	Actor Actee	91, 119, 120
(-) не показываться (-)	do not show			medium-process	Medium (Spatial-locating or enumeration) «среди»+gen	92
(отредактировать) редактировать (редактирование)	edit	Data-action EDIT	Actor Actee	dispositive-material-action	Actor Actee Manner «с помощью» +gen.	172
нарисовать рисовать рисование	draw sketch ∅+line			creative-material-action	Actor Actee	1, 26, 43, 175, 192 187, 189, 201
создать создавать (создание)	create			creative-material-action	Actor Actee	60 24, 87
указать (указывать) указание	specify enter	Simple-action SPECIFY	Actor Actee	dispositive-material-action	Actor Actee Manner «в форме» +gen	58 154, 169
выбрать (выбирать) (выбор)	choose	Select-action CHOOSE	Actor Actee Options	dispositive-material-action	Actor Actee (Spatial-locating) «в, на»+loc	85, 125
Отобразить Отображать/ся (отображение)	display	Located-action ENTER	Actor Actee Location	dispositive-material-action	Actor Actee (Spatial-locating) «в, на»+loc	87 83
Появиться (появляться) (появление)	appear			nondirected-happening	Actor (Spatial-locating) «в, на»+loc	215

Figure 8 - Table of Russian verbal and nominal complementations (material processes)

The second addition to the UM is the concept of directed-motion-processes which are material dispositive actions accompanied with a spatial locating circumstance representing a destination. The corresponding corpus data are summarized in the table in Figure 9.

Comments specific to the table in Figure 10: We have also added the concept of phase processes for verbs *повторять* (repeat), *завершить* (complete) and *продолжить* (continue). Their specific feature is that Actee is a process expressed by an infinite verb form or a nominal group representing an action. In contrast to the English *to complete the line* the Russian *закончить рисование линии* (complete drawing of the line) requires an explicit action.

Wordbasis	**English Gloss**	**UM process type**	**Corpus Occurrences**	**Corpus References**
вписать (вписывать) (-)	inscribe	directed-motion-process	Actor Actee (Destination) «в» + acc.	154
добавить добавлять (добавление)	add	directed-motion-process	Actor Actee (Destination) «в» + acc.	68, 71, 99, 102 81
преобразовать преобразовывать (преобразование)	convert	directed-motion-process	Actor Actee (Destination) «в» + acc.	174
сместить (смещать) смещение	offset	directed-motion-process	Actor Actee (Destination) «в» + acc. (Spatial-extent?) «на»	24
вернуться (возвращаться) (возвращение)	return	nondirected-motion-process	Actor (Destination) «в»+acc	56, 96
перейти (переходить) (переход)	switch	nondirected-motion-process	Actor (Destination) «в»+acc	53

Figure 9 - Table of Russian verbal and nominal complementations (directed motion processes)

Wordbasis	English Gloss	UM process type	Corpus Occurrences	Corpus References
завершить (завершать) (завершение)	complete	phase	Actor Actee (process – nominalization or infinitive)	12, 36, 95
(повторить) повторять (повторение) повторный – adj	repeat start again	phase	Actor Actee (Process, nominalization)	82 18
продолжить (продолжать) (продолжение)	resume	phase	Actor Actee (Process, infinitive or nominalization)	189

Figure 10 Table of Russian verbal and nominal complementations (phase processes)

3.3 An Analysis of the Bulgarian Data

To capture all grammatical phenomena which are used in the Bulgarian corpus and which affect the features of different lexical items we decided to group all the words presented as follows:

- verbs and some verb forms (present participle), which express the processes in the clause
- substantives and nominalized verb forms
- modifiers (adjectives, demonstrative pronouns)
- pronouns
- adverbs, prepositions and conjunctions

We discuss each of these groups in the following sections, respectively. The main idea is to compare word forms in Bulgarian corpus with the English words of corresponding group and to realize do the two (Bulgarian and English words) play the same functional role in the clauses. The KPML way of working is to extract from this functional role the proper lexical features for the word as item in the lexicon.

3.3.1 Verbs

The table in Figure 11 shows the verbs classified as "material actions" as for the Upper Model concept. The first column, **Wordbasis,** is filled with the Bulgarian verbs in alphabetical order. **English Gloss** is for the corresponding English verb from English variant of the corpus. Next two columns, **DM concept** and **DM concept slots,** show the concrete Bulgarian (and English) verbs belonging semantically to the concept of the Domain Model. Under **UM process** you can see a bit finer classification for the verb down the UM hierarchy. Next column is named **Corpus Occurrences** and shows the different combinations of complements for the verb as they occur in the corpus. The latter is explicitly marked with the number of the corpus clause, where the verb appears in its particular "surroundings". The terms used to describe the so-called "surroundings" are UM concepts: Actor, Actee, Spatial-locating,

Purpose, Symblz, Temporal-locating, Property_Ascription, Instructional, Agentive. When some of them are realized as prepositional phrase the preposition is shown in « ».

Wordbasis	English Gloss	DM concept	DM concept slots	UM process	Corpus Occurrences	Corpus References
Взема	lift up			dispositive-material-action	Actor Actee	154
Въведа	enter	Located-action ENTER	Actor Actee Location	dispositive-material-action	Actor Actee Actor Actee Spatial-locating «в» Actor Actee Purpose «за» Actor Actor Actee Symblz «под формата на»	13,16,31… 66… 107,131,176 49
Въвеждам	enter	Located-action ENTER	Actor Actee	dispositive-material-action	Actor Actee Temporal-locating	159
Върна	move back			dispositive-material-action	Actor Actee Spatial-locating «на»	183
Върна се	return			nondirected-material action	Actor Spatial-locating «в, на»	47,93,186
Добавя	add			dispositive-material-action	Actor Actee Actor Actee Spatial-locating «към»	68 78
Държа	be up			dispositive-material-action	Actor Actee	167
Завърша	complete end	Simple-action END	Actor Actee	dispositive-material-action	Actor Actee Actor	12,52,118, 173… 30
Задам	specify	Simple-action SPECIFY	Actor Actee	dispositive-material-action	Actor Actee	8,9,10, 27,28, 151…
Запиша	save record	Data-action SAVE	Actor Actee	dispositive-material-action	Actor Actee	76,166

Wordbasis	English Gloss	DM concept	DM concept slots	UM process	Corpus Occurrences	Corpus References
Затворя	close			dispositive-material-action	Actor Actee	32,121
Избера	choose	Select-action CHOOSE	Actor Actee Options	dispositive-material-action	Actor Actee Actor Actee Spatial-locating «в, на» Actor Actee Client «за»	65,67,69… 5,7,24,26,71, 82… 85
Изляза	exit			nondirected-material action	Actor Spatial-locating «от»	77
Използвам	use			dispositive-material-action	Actor Actee (Actor) Actee	3,22,36,55… 59,96,141
Изтрия	undo erase			dispositive-material-action	Actor Actee Actor Actee Time-locating «по време на»	184 15
Местя	move			dispositive-material-action	Actor Actee Spatial-locating «по»	155
Натисна	press	Simple-action PRESS	Actor Actee	dispositive-material-action	Actor Actee Actor Actee Spatial-locating «на,върху»	11,29,51… 19,106
Начертая	draw			creative-material-action	Actor Actee	34,150,163, 182…
Определя	specify enter	Simple-action SPECIFY	Actor Actee	dispositive-material-action	Actor Actee	132,135
Отворя	open	Data-action OPEN	Actor Actee	dispositive-material-action	Actor Actee	58
Отменя	undo			dispositive-mat'l-action	Actor Actee	188

Wordbasis	English Gloss	DM concept	DM concept slots	UM process	Corpus Occurrences	Corpus References
Повторя	repeat			dispositive-material-action	Actor Actee	79
Подновя	resume			dispositive-material-action	Actor Actee Actor Actee Spatial-locating «от»	169 158
Покажа се	display	Simple-action DISPLAY	Actor Actee	nondirected material action	Actor	88
Показвам се	show	Simple-action DISPLAY	Actor Actee	nondirected material action	Actor	89
Получа	list	Simple-action DISPLAY	Actor Actee	dispositive-material-action	Actor Actee=list +Property_Ascription «на»	104,105
Посоча	- (purpose «for»)			dispositive-material-action	Actor Actee	134
Поставя	put down			dispositive-material-action	Actor Actee	148
Появя се	display	Simple-action DISPLAY	Actor Actee	nondirected-material action	Actor Spatial-locating «в»	84
Появявам се	display	Simple-action DISPLAY	Actor Actee	nondirected-material action	Actor Spatial-locating «в»	80
Превключа	switch			dispositive-material-action	Actor Agentive «на»	44
Превърна	convert			dispositive-material-action	Actor Actee Agentive «в» Actor Actee Agentive «в» Instr «с»	56 138
Премахна	undo			dispositive-material-action	Actor Actee	116

Wordbasis	English Gloss	DM concept	DM concept slots	UM process	Corpus Occurrences	Corpus References
Преместя	move			nondirected-material action	Actor Actee Actor Actee Spatial-locating «в»	149 181
Преместя се	move			nondirected-material action	Actor Spatial-locating	180
Прибавям	add			dispositive-material-action	Actor Actee Agentive «към»	64
Приключа	finish	Simple-action END	Actor Actee	dispositive-material-action	Actor Actee	194
Променя	change	Simple-action EDIT	Actor Actee	dispositive-material-action	Actor Actee	110
Редактирам	edit	Data-action EDIT	Actee Action	dispositive-material-action	Actor Actee Instr «с»	54, 137
Свържа	connect			dispositive-material-action	Actor Actee Agentive «с»	14
Сложа	be down			dispositive-material-action	Actor Actee	164
Стартирам	start	Simple-action START	Actor Actee	dispositive-material-action	Actor Actee Actor Actee Instr «с»	2,18,21,95… 125
Създам	create			creative-material-action	Actor Actee	53,136
Съхраня	record	Data-action SAVE	Actor Actee	dispositive-material-action	Actor Actee Spatial-locating «в»	51
Трия	erase			dispositive-material-action	Actor Actee	192
Чертая	draw			creative-material-action	Actor Actee	156,161, 165

Figure 11 - Table of Bulgarian verbs (material processes)

Our observations and comments on the table in Figure 11 are summarized below.

Verb Aspect

We accept that the choice of the verb aspect has to be made by the Grammar according to the meaning of a particular clause. In other words, perfective and imperfective verbs bring different semantic filling. That is why we find it reasonable to organize the lexicon with different items for a perfective and imperfective verbs. For example, "Въведа" and "Въвеждам" are different lexical items and they have to be marked by adding P-ASPECT and IMP-ASPECT to their lexical features.

Reflexive Verbs

Bulgarian reflexive verbs have different meanings compared with the corresponding irreflexive verbs. This can be seen in the table in Figure 11. For example, considering UM terminology, "Върна" and "Върна се" belong to different categories:

- Върна - Dispositive-material-action involving Actor and Actee
- Върна се - Nondirected-material-action for motion, where Actor and Actee conflate.

Hence, we use different entries for them in our lexicon and to ensure the presence of reflexive pronoun "се", we need an additional feature for the reflexive verbs: REFLEXIVE. (The generation process might need explicit presence of NON-REFLECTIVE, too). The Grammar and the Text Planner are considered responsible for putting"се" in the right place-before or after the verb. In fact, in all cases in the corpus, the reflexive pronoun "се" is immediately before the verb form.

In general, Bulgarian reflexive verbs can be:

- reflexive proper (can be combined with the long form of "myself"-"себе си": *мия се, мия себе си / wash oneself*),
- reciprocal (obligatory reflexive form: *обичат се (един друг) / love each other*),
- passive,
- medial (can not combine with the long form, e.g.: *изправям се/ stand up*).

We will not make all these distinctions at this point, since the occurrence of reflective verbs is much restricted in the given AGILE corpus.

The Bulgarian Present Participle

Brackets in the column **Corpus Occurrences** for the verb *използвам/ use* are to show that in the three clauses (with numbers 59, 96 and 141) the form of the verb is NONFINITE:използвайки. There is direct semantic correspondence to English "using" (in the same three clauses). It is a morphological task to infer ing-form for any English verb. The same is with all Bulgarian imperfective verbs and their "йки"-form. We do not need a new lexical feature if IMP-ASPECT is added, as mentioned above.

1st, 2nd, 3th person; singular, plural

It is not visible in the table, but thinking about the differences between English and Bulgarian verbs, we have to mention the fact that Bulgarian morphology should deal with the problem of changes for 1st, 2nd and 3rd person singular and plural. The most frequent in the corpus are imperative clauses, which need FINITE instead of NONFINITE verb form for the English variant. The style of the text (the Text Planner) preselects the polite form: 2nd person,

plural. It is possible that some new lexical features will be needed for marking different groups of verbs according to the different ways they form their mentioned above forms (some features analogical to S-*ED feature, for example).

Problems with the DM notion "Simple-action DISPLAY"

The English verbs that instantiate this notion in the corpus are *display, show, list*.

1) In the corresponding Bulgarian clauses we find reflexive verbs: *появя се, пявявам се, покажа се, показвам се/ appear*, *show oneself*. In other words in Bulgarian text the processes are Nondirected-material action versus Dispositive-material-action in English. The problem is that the DM concept “DISPLAY” expects Actee, because it inherits the slots for the English verbs. The DM concept “DISPLAY” should be more general.
2) Bulgarian correspondence for the English verb "list" is *получавам списък/ receive a list*, which is a descriptive translation because of concrete analogy absence. The possible translations for "list" could be:

   Правя списък / make a list

   Показвам списък / show a list

   Давам списък / give a list

   Правя видим (достъпен) списък / make a list accsesible

That is why the meanings of the DM concept "Actee" are different for English and Bulgarian variants, but it's not a conjunctive case. We just need to fix ”Bulgarian” Actee = “list” and to extend/modify it further by “what are the items of the list”. Relation is Property_Ascription realized with a prepositional phrase, preposition «на» in that particular clause. Alternatively, we would have to use different DM notion for “list”.

Modal verbs

The occurrences of modal verbs in Bulgarian corpus are as follows:

Мога

17 Можете да започнете нова линия от крайната точка на последната начертана линия (You CAN start a new line at the endpoint of the last line drawn)

49 Можете да въведете тези относителни стойности под формата на @distance<angle (You CAN enter THESE relative values in the form @distance<angle)

54 можете да я редактирате с PEDIT (you CAN edit it with PEDIT)

137 можете да го редактирате с Pedit (you CAN edit it with PEDIT)

155 така ще можете да местите курсора по екрана (so that you CAN move the cursor around the screen)

165 можете да продължите да чертаете (you CAN continue drawing)

191 не можете да редактирате (you CAN’t edit them)

Трябва

50 в този случай трябва да въведете @3<100 (in this case, you WOULD enter

@3<100)

The lexical feature NONINFLECTABLE is not suitable for Bulgarian modal verbs because they have different forms depending on the person and singular versus plural. We are marking them as INFLACTABLE.

	Singular	Plural
1st Person	Мога	Можем
2nd Person	Можеш	Можете
3th Person	Може	Могат

Figure 12 Bulgarian modal verb *мога/ can*, present tense

The set of features ABILITY_AUX, POSSIBILITY_AUX etc. seems to be suitable and sufficient for Bulgarian modal clauses and for their occurrences in the corpus in particular.

3.3.2. Substantives

In general, Bulgarian nouns can be described with the same lexical features as the English ones. The features suggested for missing nouns during generation for the Initial Demonstrator texts are general and make sense for Bulgarian nouns. Some differences arise in the mechanisms and particular rules for achieving plural forms due to the rich Bulgarian morphology. We need some new features to capture the gender: MASCULINE, FEMININE, NEUTER, because plural forms (and determination, which looks as word form) depend on the gender. Nouns in Bulgarian are determined with respect to gender.

Nominalization

Only imperfective Bulgarian verbs can be nominalized. It is common in the Bulgarian language to use the nominalized form as a regular substantive with neuter gender. The nominalized forms could be determined or even used in plural (especially in a spoken language). Hence, we need different items for them in the lexicon (in contrary to inferring them from the verb) and the existing lexical property (NOMINALIZATION Original_Verb) is enough to ensure the generation, when the item is seen as process.

Corpus Occurrences	Original Verb	Gloss
81 задаване	Задавам	To specify
139 записване	Записвам	(to) record
185 изтриване	Изтривам	Erasure
171 натискане	Натискам	(you) click
107,108 подравняване	Подравнявам	To justify, justification
194 скициране	Скицирам	Sketching
57 създаване	Създавам	To create
175 триене	Трия	To erase
1,20,33 чертане	Чертая	To draw

Figure 13 Nominalized forms in the Bulgarian corpus

3.3.3. Modifiers

The main difference between Bulgarian and English adjectives in their role as modifiers is that in the Bulgarian language adjectives are sensitive to the gender of the modified noun.

Adjectives

It is expected that some new features will be necessary concerning the need of transformations for achieving feminine, neuter and plural form of adjectives, especially when the particular word has idiosyncratic forms.

The rules are: for feminine we should add “a” at the end of the word, for neuter-“o”, for plural- “?”, but by far, not each adjective follows the rules.

Demonstrative Pronouns

In the Bulgarian corpus we have following demonstrative pronoun forms:

този

50 в този случай трябва да въведете @3<100 (in THIS case, you would enter @3<100)

тези

3,36,141 като използвате един от тези методи (using one of THESE methods)

49 Можете да въведете тези относителни стойности под формата на @distance<angle (You can enter THESE relative values in the form @distance<angle)

There exists a full mapping between Bulgarian and English words in their functional role as modifiers on the one hand and as pronouns on the other hand, so we do not need new features for the corresponding lexical items.

3.3.4. Pronouns

Having lost the heavy system of cases in the Bulgarian language, we are in a situation very similar to that of English considering the cases, pronouns are concerned. The way English pronouns are presented in the lexicon (and the way they are used during generation) is suitable for the Bulgarian pronouns, too. We are going to use the same organization in the lexicon, namely different items for case forms as it is, for example, for SHE in the English lexicon, as shown in Figure 14.

```
(LEXICAL-ITEM
    :NAME                SHE ->TIA
    :SPELLING            "she"
    :SAMPLE-SENTENCE     "she can swim"
    :FEATURES
(NONE-OF-NOPOSTMODIFIERS-INDEFINITEPRONOUN-SUGGESTIVEPARTICLE-
POSSESSIVEPRONOUN
                NUMBER STEMFORM CASE NONINTERROGATIVE
PRONOUN)
    :PROPERTIES          ((CASE SUBJECT) (STEMFORM SHE) (NUMBER
SINGULAR))
    :EDITOR              "CUMMING"
    :DATE                "Thursday the tenth of October, 1985; 2:27:22 pm"
    )

 (LEXICAL-ITEM
    :NAME                HER-PRONOUN
    :SPELLING            "her"
    :SAMPLE-SENTENCE     "I like her"
    :FEATURES            (NUMBER

NONE-OF-NOPOSTMODIFIERS-INDEFINITEPRONOUN-SUGGESTIVEPARTICLE-
POSSESSIVEPRONOUN
                STEMFORM CASE NONINTERROGATIVE PRONOUN)
    :PROPERTIES          ((CASE OBJECT) (STEMFORM SHE) (NUMBER
SINGULAR))
    :EDITOR              "CUMMING"
    :DATE                "Thursday the tenth of October, 1985; 2:32:39 pm"
    )
```

Figure 14 Lexical entries for pronouns

Whether all the presented features are necessary depends on the particular situations while generating our particular examples in the corpus. The occurrences of pronouns in the Bulgarian variant of the corpus are as follows

Вие

151 която ВИЕ сте му задали (YOU specified)

Го

137 можете да ГО редактирате с Pedit (you can edit IT with PEDIT)

138 или да ГО превърнете в самостоятелни сегменти от линии с (EXPLORE or convert IT to individual line segments with EXPLODE)

183 и след това ГО върнете по линията на мястото (and then move IT back as far along the line as)

Я

54 можете да Я редактирате с PEDIT (you can edit IT with PEDIT)

56 за да Я превърнете в самостоятелни линии и дъги (to convert IT to individual line and arc segments)

123 за да Я завършите (to end IT)

	Nominative Case	Objective Case
1st Person Singular	Аз / I	Мене / me
2nd Person Singular	Ти / you	Тебе / you
3th Person Singular masculine feminine neuter	 Той / he Тя / she То / it	 Него / him Нея / her Него / it
1st Person Plural	Ние / we	Нас / us
2nd Person Plural	Вие / you	Вас / you
3th Person Plural	Те / they	Тях / them

Figure 15 Personal pronouns: Bulgarian - English correspondence

Talking about the lexicon being exhaustive (and trying to capture pronoun occurrences in our corpus), we need one additional item per each of the rows in the table in Figure 12, namely for the short form in the objective case. This additional form should be distinguished from a long objective form by a new feature: SHORTFORM.

	Nominative Case	Objective Case	Objective Case SHORTFORM
1st Person Singular	Аз / I	Мене / me	Ме
2nd Person Singular	Ти / you	Тебе / you	Те
3th Person Singular masculine feminine neuter	 Той / he Тя / she То / it	 Него / him Нея / her Него / it	 Го Я Го
1st Person Plural	Ние / we	Нас / us	Ни
2nd Person Plural	Вие / you	Вас / you	Ви
3th Person Plural	Те / they	Тях / them	Ги

Figure 16 Personal pronouns: Bulgarian and English correspondences (extended)

An interesting phenomenon in the Bulgarian language is clitic-doubling (using together and immediately one after another long and short objective personal pronoun: него го, нея я, нас ни…). However, clitic doubling does not occur in the AGILE corpus, because of the formal style of the CAD_CAM documentation, so we will not go into more detail on this issue.

3.3.5. Adverbs, Prepositions and Conjunctions

There are no significant differences in the functional role of these three categories of words mapping Bulgarian to English words. "One to one" mapping gives us the possibility to use the definitions of the existing English lexical items.

3.4 Overall Conclusions Based on the Empirical Data

From the observations about the gathered empirical data we can draw various conclusions:

1.The variety of complementation occurrences in the corpus is a subset of the theoretically possible valency frames for verbs of a particular language.

2. Complements may have to be realized using a specific morphological case, depending on the *individual* verb (Czech, Russian).

These conclusions lead to a set of desiderata that will direct the way in which we will model lexical resources in KPML:

- It should be possible to specify for a verb (or noun) how its realization may deviate from the way its underlying concept has been defined. For example, given the concept APPEAR, which has slots Actor and Location that are both obligatory, we should be able to allow the Czech verb *objevit se* to be realized without Location-complement[21] - in contrast to the Russian verb *pojavitj sja* that must be realized with a Location-complement. This desideratum concerns the conclusion 1 above.

- It should be possible to specify in a lexical entry for a verb, how its complements are to be realized. This desideratum concerns conclusion number 2.

[21] Even though the valency frame of the verb *does* include the LOCATION as an obligatory free modifier (mirroring the concept).

4. Modelling Lexical Resources for Slavic Languages in KPML

4.1 Basic Form of Lexical Resources in KPML

Let us consider again the basic form of a lexical entry in KPML, as we already saw it earlier:

```
(lexical-item
      :NAME name-of-the-item
      :SPELLING ``spelling of the item''
      :SAMPLE-SENTENCE ``a sample sentence using the item''
      :FEATURES (lexical-feature-1 .... lexical-feature-n)
      :PROPERTIES (property-1 ... property-m)
      :EDITOR ``name of the editor''
)
```

Figure 17 - Abstract form of a lexical entry in KPML

The :NAME field in a lexical entry specifies a label by which we can refer to the specific entry by a :LEX statement in an SPL. Although the value of the :NAME field can be anything, we will adhere to the convention here that the :NAME of a lexical entry is the base form of the word that the entry defines.

The :FEATURES field characterizes the lexico-grammatical nature of the word. As we already said above, when we use the grammar to generate a sentence, we construct a representation of the surface form in terms of *constraints* that this form is to obey. Such is achieved by spelling out the grammatical features that altogether would make the sentence to exhibit the meaning as given in the SPL, and then by realizing the grammatical features in terms of grammatical functions, constituents, and morphological form.

Essentially there are two ways, then, that lexical features can function in the process of generation. Either the SPL specifies the lexical entries we should use, or it does not. In the latter case, it is up to the grammar to associate specific lexical features to functions, which is possible by means of the *classify* and *outclassify* realization operations. If, on the other hand, lexical items are given, then the lexical features will act as constraints that are to be taken into account when deciding for one realization or another. That is to say, if there are alternative realization statements but only one realization statement would connect the proper lexical features to grammatical functions, then only that realization statement could be chosen.

Finally, the :PROPERTIES-list can be used to describe idiosyncratic exceptions to the general morphological realization of words that belong to the same wordclass as the word for which the entry is defined. For example, for the verb “feed” we can define the irregular PASTFORM “fed” - thus ensuring that we will not obtain a sentence in which the past tense of “feed” is realized as “feeded”. The purposes of the remaining fields in a lexical entry are self-evident.

4.2 Encoding the Desiderata as Lexical Features

We propose to implement the desiderata mentioned at the end of section 3 by means of

lexical features. This is one option out of two, essentially.

One option would be to conceive of the realization of a complement in a specific (morphological) case as being associated to an aspect of *meaning*. That is to say, depending on the meaning to be expressed, we would make a decision for realizing a complement as, for example, either a nominal group in dative case or as a nominal group in accusative case. The consequence of adhering to this view would be that we would first of all define appropriate *grammatical features* that would be used to characterize a structure as a whole, depending on the meaning to be expressed. In the grammar, we would then define systems that would relate these grammatical features to realization statements.

There have been various authors that have proposed to view morphological inflection as expressing propositions, and to treat them as such. Also, some have argued that the choice for realizing a grammatical *function* as a nominal group in a particular case does not depend on the verb the grammatical function is a complement of, but is rather regards concerns surrounding the textual metafunction.

Contrary to these (rare) views we would rather follow the general belief that requirements on morphological form of a verb's or a noun's complements are purely a matter of *lexicogrammar*, not of meaning. Therefore, we would like to propose to implement the desiderata as *lexical features*, which can be specified in individual lexical entries, and which will appear in the relevant systems binding functions and lexical features in their realization statements. Which systems *are* relevant, theoretically, is discussed in the next section; in the remainder of this section we will discuss which new lexical features we would like to propose.

Definition of Additional Lexical Features

On the basis of the observations from the corpus data, we introduce the following lexical features. These lexical features implement the second desideratum. Observe that we define the lexical features as *templates* here, in the sense that for the [X] we can fill in any (sensible) grammatical function.

- DATIVE-[X]-IN-ACTIVE-VOICE. The grammatical function *X*, in its role as a complement of a verb (or a noun), needs to be in dative case when it occurs in a clause that is in active voice. For example, the Czech verb ..... requires its Goal to be in dative case. Hence, in the :FEATURES-list of the verb's lexical entry we should specify DATIVE-GOAL-IN-ACTIVE-VOICE in order to impose the constraint.
- ACCUSATIVE-[X]-IN-ACTIVE-VOICE. Similar to the previous lexical feature, only this time the grammatical function *X* should be in accusative case.
- INFINITIVE-[X]. This lexical feature expresses that the grammatical function *X* should be an infinitival construction. Voice does not matter in this case.

The following feature is defined in order to deal with the first desideratum.

- OBLIGATORY-LOCATION-DELETABLE.
- OBLIGATORY-DIRECTION-DELETABLE.

4.3 Systemic Concerns

First, we should recall that a grammar in KPML is conceived of as a network of systems,

whereby the network's organization follows that of *rank scale*. At the top of the rank scale, we find the clause and various kinds of groups, whereas at the bottom of the scale we find morphological inflexion[22]. Accordingly, when traversing the network we first encounter systems dealing with the realization of clauses, and then verbal groups and nominal groups, whereas only in the end we come to systems that deal with imposing constraints on the morphological form of specific expressions.

It is then important to realize that the desiderata and the lexical features achieving them are not all related to the same rank. Rather, the first desideratum and the relevant lexical features indicating the possibility to drop ("delete") an obligatory complement are (to be) located at the rank of *groups*. The systems implementing that rank in the grammar deal with the organization of verbs and their complementations.

On the other hand, the second desideratum and its corresponding lexical features deal with *words and their morphological form*, and are thus located at a much lower rank. Systems imposing constraints on morphological form are usually invoked much later in a traversal through the network than the systems dealing with (verbal) groups.

4.4 Prototypical Description of Implementation

Here we provide a discussion of changes in, and additions to, the Nigel grammar to reflect sensitivity to the lexical features introduced above, and to reflect some more general aspects of sentences in Slavic languages. We present here *prototypical* implementation of systems.

Systems dealing with complementations and their morphological form

The purpose of these systems is multifold. First of all, we need to add systems to ensure that the resulting grammar becomes sensitive to lexical features that specify what form a specific complement of a verb or noun needs to have. Examples of suchlike lexical features are DATIVE-GOAL-IN-ACTIVE-VOICE and ACCUSATIVE-GOAL-IN-ACTIVE-VOICE defined in the preceding section.

These systems are -essentially- to deal with imposing the proper constraints on the morphological realization of a constituent. Naturally we can place this in a larger context, since such systems may be needed for slightly different purposes as well: namely, in Slavic languages *prepositions* always go with one or more cases which are specific to the preposition used and the meaning to be expressed. Two interesting issues can subsequently be raised:

1. Where should we place the systems that relate (by means of a *classify* realization operation) the relevant lexical feature(s) to grammatical functions (e.g., DATIVE-GOAL-IN-ACTIVE-VOICE to a Goal once the latter gets inserted)?, and
2. How do we generate proper word forms, based on the imposed morphological constraints?

To begin with the first point, these systems should be added to the network after the systems introducing the relevant grammatical function. We work out the case for a *Goal*.

In the Nigel grammar[23], a *Goal* is inserted in the following way. In case we are dealing

---

[22] At least, for morphologically rich languages.

[23] Recall that in AGILE, we take the Nigel grammar, which was originally developed for English, and try to revise it such that it works for a specific Slavic language.

with an effective verb, and a *Material* has been introduced by transitivity, the grammatical feature *effective-material* will be inserted. Provided that also a *Medium* (grammatical function) has been inserted as a *Direct Complement* (i.e. as grammatical object), the grammar will introduce a *Goal,* conflate it with the *Medium*, and preselect the Goal to be a nominal group. To verify this process, one can for example generate sentence 10 from the 2A-example set that has been used for the initial generator in AGILE (WP 7.1).

Let's have a look at the introduction of a *Goal* in more detail. It happens in a system aptly called **Goal-Insert-Conflate**, which is defined as follows:

```
(GATE
      :NAME   GOAL-INSERT-CONFLATE
      :INPUTS   (AND EFFECTIVE-MATERIAL MEDIUM-INSERTED)
      :OUTPUTS   ((1.0 GOAL
        (INSERT GOAL)
        (CONFLATE GOAL MEDIUM)
        (PRESELECT GOAL NOMINAL-GROUP)))
      :CHOOSER   GOAL-INSERT-CONFLATE-CHOOSER
      :REGION   NONRELATIONALTRANSITIVITY
      :METAFUNCTION   EXPERIENTIAL
  )
```

Figure 18 - System for inserting a Goal

whereby the chooser associated to the system is simply defined as

```
  (CHOOSER
     :NAME   GOAL-INSERT-CONFLATE-CHOOSER
     :DEFINITION   ((CHOOSE GOAL))
  )
```

Figure 19 - Chooser for GOAL-INSERT-CONFLATE

The changes we need to make to the system (and the chooser) are the ensuing ones. To begin with, we need to introduce more distinctive *grammatical* features in the system. Because we are specifying prototypes here, we can assume for the moment that we need only distinguish between a Goal to be realized in dative case or in accusative case.

In the revised version of the system we employ corresponding grammatical features *Goal-Dative* and *Goal-Accusative*, which will have different associated sets of realization statements. These sets will impose constraints on the Goal to be realized as a nominal group in dative case or in accusative case, respectively. Observe that we explicitly link the

morphological realization of the Goal to requirements that (are assumed to) appear in lexical entries. Of course we only specify dative or accusative case - additional constraints on plurality or singularity of the Goal, as well as the gender, are introduced by other systems (under the **Nerb** system).

```
(GATE
      :NAME   GOAL-INSERT-CONFLATE
      :INPUTS   (AND EFFECTIVE-MATERIAL MEDIUM-INSERTED)
      :OUTPUTS   (
        (0.3 GOAL-DATIVE
            (INSERT GOAL)
            (CONFLATE GOAL MEDIUM)
            (CLASSIFY GOAL NOUN)
            (CLASSIFY GOAL DATIVE-GOAL-IN-ACTIVE-VOICE)
            (INFLECTIFY GOAL DATIVE-FORM)
         )
        (0.7 GOAL-ACCUSATIVE
            (INSERT GOAL)
            (CONFLATE GOAL MEDIUM)
            (CLASSIFY GOAL NOUN)
            (CLASSIFY GOAL ACCUSATIVE-GOAL-IN-ACTIVE-VOICE)
            (INFLECTIFY GOAL ACCUSATIVE-FORM)
      )
      :CHOOSER   GOAL-INSERT-CONFLATE-CHOOSER
      :REGION   NONRELATIONALTRANSITIVITY
      :METAFUNCTION   EXPERIENTIAL
  )
```

Figure 20 - Prototype system for different morphological realizations of a Goal

The revised version of the chooser is:

```
  (CHOOSER
     :NAME   GOAL-INSERT-CONFLATE-CHOOSER
     :DEFINITION   ((DEFAULTCHOOSE GOAL-ACCUSATIVE))
```

Figure 21 - Revised chooser for GOAL-INSERT-CONFLATE

The chooser opts by default for the GOAL-ACCUSATIVE case, which is the usual case for

goals in active voice (in Slavic languages), unless it is constrained otherwise. This may be language-dependent, of course.

The association of the lexical feature DATIVE-GOAL-IN-ACTIVE-VOICE to the grammatical function Goal has the following impact on the way the surface form of sentence will get filled in with words from the lexicon. In the typical case that we are generating a sentence from an SPL that does not contain any :LEX statements (i.e., does not restrict us to use specific lexical entries), then lexification will be able select only those lexical entries as possible realizations which obey all the constraints and include the selected lexical feature (i.e. either DATIVE-GOAL-IN-ACTIVE-VOICE or ACCUSATIVE-GOAL-IN-ACTIVE-VOICE). On the other hand, if the SPL *does* specify which lexical entry to use, then the choice between GOAL-ACCUSATIVE and GOAL-DATIVE will depend on what lexical feature the lexical entry's :FEATURES-list contains. In other words, the choice is constrained to that option which is satisfiable with the constraints imposed by the lexical entry's :FEATURES.

Finally, the way the proper wordforms are generated in AGILE is as follows. Once the proper constraints on case have been introduced, for example by means of systems like the one above, and additional constraints have been introduced on number (via the **Singular-Form** or **Plural Form** systems under the **Nerb** system) and gender (to be introduced in the Nigel grammar), then an external morphological component will take care of the generation of the proper forms. We described the use of morphological components in the deliverables MORPH1 (WP 4.1)[24].

Systems dealing with possible deletion of obligatory complementations

The corpus data already showed us that languages may differ in whether it is possible to delete (omit) the realization of a complementation, even though the complementation would be obligatory. In Russian, for example, it seems that the verb "to appear" can not be realized without a constituent realizing the grammatical function *Location*. On the contrary, in Czech such *is* possible, in case the location is unspecific and thematized. For example, we always assume (in the context of AGILE) that when a dialogue box or a toolbar appears, it always appears on a *computer screen* (which seems a reasonable assumption). Therefore, in Czech we would not normally realize the location.

We take the following approach to modelling this phenomenon. Whether or not a Location should be realized is a choice to be made *after* the choice has been that the Location should be expressed (as opposed to not expressed)[25]. We can translate this in terms of the Nigel grammar as follows.

Under **Transitivity-unit** there is the possibility to either express a Location (**Spatial-Locative)** or not (**No-Spatial-Locative**). It is the system **Spatial-Locative** which inserts a grammatical function *Space-Locative*, which subsequently will get realized further in terms of a *MiniRange* (nominal) and a *MiniProcess* (preposition). Instead of unconditionally inserting a *Space-Locative* in the **Spatial-Locative** system, we could make the insertion dependent on the context.

---

[24] In terms of a KPML/SFL-style grammar, the consequence is that we will not define any systems under the **Morphemes** system, which implements the lowest **Rank** system.

[25] There are various arguments for taking this perspective. For one, this would still enable one to express a Location in an A-box or an SPL. Another, related argument could be that this way you are modelling a choice whether or not Location could be *elided*.

We can do so by distinguishing two grammatical features, *Location-Realized* and *Location-Not-Realized*. The *Location-Realized* feature has an associated realization operation which inserts the grammatical function Spatial-Locative. The *Location-Not-Realized*, on the other hand, should insert of a *Spatial-Locative* and then preselect the *Spatial-Locative* to be realized as an empty string (empty category). What is important is that *Location-Not-Realized* should relate the inserted (though empty) *Spatial-Locative* function to the lexical feature OBLIGATORY-LOCATION-DELETABLE by means of a "classify" realization operation.

The most interesting part of this system is the chooser. It is there that we decide whether or not we can delete (omit) the Location:

```
(CHOOSER
   :NAME   SPATIAL-LOCATIVE-CHOOSER
   :DEFINITION   (
       (ASK (DELETABLE-LOCATION-Q DELETABILITY)
          (NONDELETABLE
              (CHOOSE LOCATION-REALIZED))
          (DELETABLE
              (ASK (SPECIFIC-Q SPECIFICITY)
                  (SPECIFIC
                      (CHOOSE LOCATION-REALIZED))
                  (NONSPECIFIC
                      (ASK (THEMATIC-Q THEMATICITY)
                          (NONTHEME
                            (CHOOSE LOCATION-REALIZED))
                          (THEME
                            (CHOOSE LOCATION-NOT-REALIZED)
                      )
              )
       )
)
```

Figure 22 - Location chooser

This way, a Location may be deleted if and only if it is deletable, unspecific, and part of the theme. Clearly, text structuring (WP 5.2) will have an important impact on this issue. Note that the inquiries concerning thematicity and specificity concern meaning, whereas the inquiry concerning deletability can be formulated as a language-dependent inquiry with a default reply (for Czech, "deletable"; for Russian, "nondeletable").

A Brief Sideremark

Finally, an important role in the realization of a Space Locative appears to be the kind of

process it is a part of. As a side remark, we would like to point out the following discrepancy between the way the Nigel realizes *circumstantial relations* for English, and how such can be done in Slavic languages. In Nigel, more specific grammatical features like *Minor-Process-Type*, are all inserted by systems that are placed in the network under the *Prepositional Phrase* system. Put slightly differently, Nigel will always realize a *circumstantial relation* (temporal, spatial, etcetera) as a prepositional phrase, and will subsequently determine the form of the prepositional phrase in more detail by chosing one more specific kind or another.

However, in Slavic languages we can realize a *circumstantial relation* either as a nominal phrase or as a prepositional phrase. Thus, instead of placing the choice for a particular kind ***after*** (or under) the prepositional phrase system, we would rather place this choice directly under the system dealing with *circumstantial relation* and leave the realization of a specific kind as a nominal phrase or as a prepositional phrase to a subsequent traversal of the grammar.

This sideremark has important consequences for the way prepositional groups and nominal groups are to be modelled in AGILE (WP 6, “Linguistic Specification”)

References

Bateman, John, *KPML Development Environment: Multilingual resource development and sentence generation*. GMD-Studien Nr. 304, GMD Forschungszentrum Informationstechnik GMBH: Darmstadt Germany, December 1996

Bateman, John, Kasper, Robert, Moore, Joanna, and Whitney, Richard, "A General Organization of Knowledge for Natural Language Processing: The Penman Upper Model". Technical Report, University of Southern California/Information Sciences Institute: Marina Del Rey CA, 1990.

Hajičová, Eva. *Agent or Actor/Bearer?* In: Theoretical Linguistics, 6, pp. 173-190, 1979.

Hajičová, Eva. *Remarks on the meanings of cases*. In: PSML 8, pp. 149-157, 1983

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

Panevová, Jarmila. *On verbal frames in Functional Generative Description*, Part One, In: PBML 22, pp. 3-40, Part Two, In: PBML 23, pp. 17-52. 1974-1975.

Panevová, Jarmila. *Verbal frames revisited*. In: PBML 28, pp. 55-72. 1977

Panevová, Jarmila. *Inner participants and free adverbials*, In: PSML 6, pp. 227-254, 1978;

Pollard, Carl, and Sag, Ivan A., *Head-Driven Phrase Structure Grammar*. The University of Chicago Press: Chicago IL, 1994 (Studies in Contemporary Linguistics)

Sgall, Petr, Hajičová, Eva, and Panevová, Jarmila, *The Meaning of the Sentence in Its Semantic and Pragmatic Aspect*. D. Reidel Publishers: Dordrecht, the Netherlands, 1986.