БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ . BULGARIAN ACADEMY OF SCIENCES

ПРОБЛЕМИ НА ТЕХНИЧЕСКАТА КИБЕРНЕТИКА И РОБОТИКАТА, 52
PROBLEMS OF ENGINEERING CYBERNETICS AND ROBOTICS, 52

София . 2001 . Sofia

Multilingual Natural Language Generation (Experience from AGILE Project)[1]

Kamenka Staykova

Institute of Information Technologies, 1113 Sofia, email: staykova@iinf.bas.bg

1. Introduction

Multilingual Natural Language Generation is an interesting and challenging field of Natural Language Processing. Automatic generation of texts in natural language could be viewed as a final part of automated translation process from one language to another. Alternative approach is given the chance with development of modern Natural Language Processing technologies, which concentrate the researchers attention on Natural Language Generation as more self-dependent area. Main idea in this respect is the claim that generation of particular text in particular language is a surface realization of particular semantic, which is communicated at the moment in the chosen language. If we have an access to (or description of) that semantic and if we use lexico-grammatical resource, which generates in another natural language we could produce a surface realization of the same semantic in the second natural language.

The theoretical base for such a belief is built by functional view to the natural languages. Halliday's functional analysis of English, Chinese and some other languages is fundamental in this respect. "Introduction to Functional Grammar" describes the basic principles of Systemic Functional Linguistics extracted by analysis of English. Beginning with the most general meta-functions of language (ideational and interpersonal) Systemic Functional Linguistics introduces the idea of communicative functions of language as common for all natural languages.

One of the most successful implementations of Hallidey's theory is the environment KPML, which inherits PENMAN system for Natural Language Generation. KPML gives the researchers basic tools for multilingual sentence generation and resource development. Multilingual generation within KPML starts from common semantics for all languages and traverses the lexico-grammatical resource of each

[1] This work is partially supported by EU INCO-COPERNICUS project PL 961104 "Automatic Generation of Instructional Texts in the Languages of Eastern Europe" and by theme of IIT, BAS, No 010037 "Methods for Knowledge Representation and Knowledge Engineering in Modern Information Technologies".

language to compose the syntactic structure of the sentence(s), which express the semantic source.

AGILE project gave the involved researchers opportunity to examine and improve the method of fast prototyping the lexico-grammatical resources for multilingual generation within KPML. For the chosen languages of Eastern Europe - Bulgarian, Czech and Russian computational functional lexico-grammatical resources hadn't existed before the project. For three years during the project were developed and tested resources for natural language generation in Czech, Bulgarian and Russian on the base of English Grammar NIGEL, which grammar is a result of more than 15 years of work. Of course, the lexico-grammatical resources for the three Slavic languages are not "large-scaled" as NIGEL is, but they could generate significant amount of clause variations (and even texts) from the chosen domain of CAD-CAM instructions.

2. Systemic Functional Linguistics

2.1. Functional view to natural languages

The main starting question of Functional Linguistics is the question "How the language is used?" Halliday describes the principles of "Functional idea" in "Introduction to Functional Grammar" (IFG) [1][2]:

"(1) Every text – that is, everything that is said or written – unfolds in some context of use;

(2) All languages are organized around two main kinds of meaning, the 'ideational' or reflective, and the "interpersonal" or active.

(3) Each element in a language is explained by reference to its function in the total linguistic system."

"Ideational" and "Interpersonal" functions of language "...underlie all uses of language: – to understand the environment (ideational, reflective) and to act on the others in it (interpersonal, active). They are called 'metafunctions' in the terminology of Functional Linguistics. Combined with these is a third metafunctional component, the "textual", which breathes relevance into the other two."

"The table below introduces the technical names for the metafunctions, matches them up with the different statuses of the clause... It will be seen that there is a fourth metafunctional heading which does not show up in the "clause" column, because it is not embodied in the clause but in the clause complex."

Table 1. Metafunctions and their reflexes in the grammar [3]

Metafunction	Definition (kind of meaning)	Corresponding status of clause
experiential	construeing a model of experience	clause as representation
interpersonal	enacting social relationships	clause as exchange
textual	creating relevance to context	clause as message
logical	constructing logical relations	–

[2] All quoted texts in section 2 are taken from [1], p. 36.

[3] The table is taken from [1], p. 36.

2.2. Systemic theory

Systemic theory used by Halliday to describe the English Functional Grammar is "...largely based on Firth's system-structure theory, but derives more abstract principles from Hjelmslev and owes many ideas to the Prague school. The organizing concept is "system", which is used essentially in Firth's sense of a functional paradigm but developed into the formal construct of a "system network".

The system network is a theory about language as a resource for making meaning. Each system in the network represents a choice... The system includes (1) the entry condition (where the choice is made), (2) the set of possible options and (3) the "realizations" (what is to be done- that is what are the structural consequences of each of the options)."

2.3. Clause in the center of attention

"Sentence and word are the two grammatical units that are recognized in our folk linguistics; and this incorporates a piece of good common sense."

"Below the sentence, the typical relationship is a constructional one, of parts into wholes. In functional grammar this means an organic configuration of elements, each having its own particular functions with respect to the whole.

Above the sentence the position is reversed. Here the non-constructional forms of organization take over and become the norm, while only in certain cases, particular kinds of texts, are there recognizible units like the structural units lower down.

The sentence, then, does constitute a significant border post, which is why writing systems are sensitive to it and mark it off."

Bulgarian, Czech and Russian belong to Slavic languages, but they like English are "Europian languages" and clause has the same status in their linguistic systems. Analyzing Bulgarian, Czech and Russian we could follow the structure used by Halliday in IFG for English. The rest of Section 2 is a theoretical generalization valid for the four languages.

2.4. Clause as message. Theme and Rheme

"In English, as in many other languages, the clause is organized as a message by having a special status assigned to one part of it. One element in the clause is enunciated as the theme; these then combines with the remainder so that the two parts together constitute a message." ([1], p. 38).

In the four languages Theme in a clause is put first, then the elements of Rheme function follow. Word order is substantial when discribing Theme and Rheme functions in Slavic languages. There are a lot of similarities between English from one hand and Bulgarian, Czech, Russian from other, but the word order in Slavic languages is much more free.

Table 2. Theme-Rheme structure[4]

Theme	Rheme
The duke	has given that teapot to my aunt.
To my aunt	has been given that teapot.
That teapot	the duke has given to my aunt.

[4] Table 2 is from [1], p.38.

2.5. Clause as exchange. Mood element (Subject, Finite) and Residue

"In the act of speaking, the speaker adopts for himself a particular speech role, and in so doing assigns to the listener a complementary role which he wishes him to adopt in his turn. For example, in asking a question, a speaker is taking on the role of seeker of information and requiring the listener to take on the role of supplier of the information demanded.

The most fundamental types of speech role, which lie behind all the more specific types that we may eventually be able to recognize, are just two: (1) giving, and (2) demanding)." [1] Page 68.

Table 3. Speech functions and responses [5]

Function	Initiation	Expected response	Discretionary alternative
Give goods&services	offer	acceptance	rejection
Demand goods&services	command	undertaking	refusal
Give Information	statement	acknowledgement	contradiction
Demand Information	question	answer	disclaimer

Presented principles in giving and demanding information work in Slavic languages the same way as in English.

In our particular work on AGILE project we had to deal with the role of Hearer among all roles (Speaker, Listener, Hearer) associated to the speech acts (see Section 4).

2.6. Clause as representation

"Language enables human beings to build a mental picture of reality, to make sense of what goes on around them and inside them. Here again the clause plays the central role, because it embodies a general principle for modelling experience – namely, the principle that reality is made up of PROCESSES...

What is the status of a process, as set upin the grammar of the clause? The framework is very simple... A process consists, in principle, of three components:

(i) the process itself;
(ii) participants in the process;
(iii) circumstances associated with the process.

This tripartite interpretation of processes is what lies behind the grammatical distinction of word classes into verbs, nouns, and the rest, a pattern that in some form or other is probably universal among human languages." We can express this as in Table 4.

Table 4. Typical functions of group and phrase clauses[6]

Type of element	Typically realized by
(i) process	Verbal group
(ii) participant	Nominal group
(iii) circumstance	Adverbial group or Prepositional phrase

[5] Table 3 is taken from [1], p.69, Table 4(1).

[6] Table 4 is taken from [1], p.109, Table 5(1).

Giving evidences from English grammar, Halliday answers the question "What are the different types of processes, as construed by the transitivity system in the grammar?". Table 5 below shows the summary of "the picture derived from English". Doing the same type of analysis for Bulgarian, Czech and Russian we can claim that in these languages Types of Processes are the same and Participants are specified the same way in Slavic languages, although some differences exist.

Table 5. Process types, their meanings and key participants[7]

Process type	Category meaning	Participant
Material	"doing"	Actor, Goal
Action	"doing"	
Event	"happening"	
Behaveoral	"behaving"	Behaver
Mental	"sensing"	Senser, Phenomenon
Perception	"seeing"	
Affection	"feeling"	
Cognition	"thinking"	
Verbal	"saying"	Sayer, Target
Relational	"being"	
Attribution	"attributing"	Carrier, Attribute
Identification	"identifying"	Identified, Identifier; Token, Value
Existential	"existing"	Existent

3. KPML

KPML is an environment for multilingual linguistic resource development and sentence generation. It enables researchers to test an existing lexico-grammatical resource, to trace the generation process, to make changes in the grammatical system network, to work with sets of clauses and so on. The interface meets the expectations of people working with modern software products. KPML Development Environment, documentation, tutorials and some grammar resources are available free from KPML site: http://purl.org/net/kpml

"KPML Resource for Natural Language Generation" – what is it like?

Lexico-grammatical resource for NLG within KPML contains lexicon(s) with functional features of the items, computational functional grammar (system network) for the language, Domain Model of particular field of generation and Upper Model, which gives the generalized notions in the world of generation. The first two "modules' are the most language dependent ones, so they are called together 'lexico-grammar' of particular language. Domain Model gives the context of generation and describes the notions of that context as 'concepts', so that they could be related to the lexical items in particular language. Upper Model gives the most generalized hierarchy of the abstract notions at all. It is an ontology extracted from the language

[7] Table 5 is taken from [1], Table 5(6), p.143.

(first it was English, then German, Dutch, Italian were analyzed and the claim of the authors [9] is that the (current) Generalized Upper Model could support natural language generation in many languages).

–**Ontology:** When using natural language to communicate people describe their reality by concepts. It is normal to be expected that when people use one and the same language they will grasp the concepts in similar ways and build one and the same Ontology.

The idea to extract ontology from Natural Language is realized by researchers in the field of Natural Language Processing ([8], [9], [10]) in their attempt to find a general organization of knowledge, which allows "significantly simplifying the interface between domain-specific knowledge and general linguistic resources". The hierarchy called UPPER MODEL and used in the AGILE project is available on KPML site.

–**Domain Model:** Process of text generation needs clear definitions in the particular field. In a sense Domain Model suggests more clear connection between the concept in its general meaning (from Upper Model) and particular lexical realization of the concept (from the Lexicon). Defined in a particular domain some ambiguous lexical items are resolved; for example in the domain of software instructions DM::MOUSE is no animal but a device for data input of computer configurations.

–**Computational Systemic Functional Grammar:** The generation procedure is that of traversing grammar network (system network) of functional alternatives. The key possition in a computational grammar network of particular language takes the system RANK. During the process of clause generation the RANK system could be entered several times in the attempt to be completed different levels/ranks of generated structure (clauses, groups/ phrases, words, morphemes).

In the particular work on the AGILE project only the ranks of clauses (simple clauses and clause complexes) and group/ phrases rank (nominal groups, verbal groups, prepositional phrases) were developed for Slavic languages. The possibility to use external modules (producing morphological word forms) is appreciated [12], because such an organization of the work allows concentrating the efforts in particular direction and at the same time achieving the surface realization of target texts, which is very difficult in the natural language generation area.

–**Lexicons:** Lexical items of this kind of resource for natural language generation are described by their functional (grammatical or morphological) features and listed in Lexicons. Some features of the lexical items serve to the process of generation and are crucial for the particular choice of a lexical item in particular clause under generation. Other type of features (namely morphological) are used in the process of producing particular word form after the choice of the lexical item.

4. AGILE Project

The aim of AGILE project was to develop a multilingual authoring tool that enables experts in writing software instructions to compose and produce software manuals in Czech, Bulgarian and Russian without any linguistic training in the three Eastern European languages or experience in knowledge representation languages. The specific application domain chosen for the prototype document generator was CAD-CAM. Target instructional texts are specified in conceptual, abstract way by combining Domain Model notions, which don't belong to particular natural language. Automatically generated AGILE target texts are in different styles used in technical

documentation (personal and impersonal) and in different types (procedures, quick reference, functional descriptions).

One of the primary goals of the AGILE project has been the development of lexico-grammatical resources suitable for multilingual generation:

"The overall goal of AGILE is to make available a generic set of tools and linguistic resources for generating Czech, Bulgarian and Russian: no such resources are currently available." (Agile, 1997).

These formed the main contribution of the project as the developed lexico-grammatical resources are the only existing functional Natural Language Generation grammars for these languages.

4.1. Resources for Multilingual Generation within AGILE prototype

–**Ontology:** Current version of UPPER MODEL (1998) was sufficient to support concepts specification in AGILE Domain Model and generation in the tree Slavic languages without any changes.

–**Domain Model:** The Domain Model built for AGILE project contains 212 concepts. Some of them define the structure of the target texts, for example the concepts PROCEDURE and METHOD-LIST from the Table 6 below. Some of Domain Model concepts are Processes or Things from the particular context of CAD/CAM domain, for example the concept DRAW is defined as Process, the concepts LINE-OBJECT and LINE are defined as Objects (see the Table 6 below).

–**Lexico-grammatical resources:** For each of the three languages was built a grammatical network of functional alternatives (systemic network) on the base of theoretical systemic functional analysis of the language. In fact, the need of fast resources development forced the researchers to use an approach-combination of a system-oriented method of grammar development and an instance-oriented one. 'System-oriented' means building up a computational resource with a view to the whole language system; 'instance-oriented' means being guided by a register analysis. Creating resources, which generate in Bulgarian, Czech and Russian wouldn't be possible within the given time without the English grammar NIGEL taken as a ground resource and a base of cross-linguistic comparisons. The method of resource development on the base of already existing resource has been claimed to be effective [3, 4] and our experiences from AGILE proved it [5].

Table 6. Definitions of concepts in the Domain Model

```
(define-concept PROCEDURE (INSTRUCTION-SCHEME)
  ((GOAL :type USER-ACTION)
  (METHODS :type METHOD-LIST :optional T)
  (SIDE-EFFECT :type USER-EVENT :optional T)))
(define-concept METHOD-LIST (INSTRUCTION-SCHEME
                                          CADCAM-LIST)
  ((FIRST :type METHOD*)
  (REST :type METHOD-LIST :optional T)))
(define-concept SIMPLE-ACTION (USER-ACTION-DIRECTED))

(define-concept DRAW (SIMPLE-ACTION)
  ((ACTEE :type GRAPHICAL-ACTEE)))

(define-concept LINE-OBJECT (UNIQUE-GRAPHICAL-OBJECT))

(define-concept LINE (LINE-OBJECT))
```

Lexicons with functional features of lexical items were organised for each of the languages as a part of particular computational resource. For example, the Bulgarian lexicon, which supports all AGILE target texts generation, contains 312 lexical items. Each lexical item has NAME, SPELLING and FEATURES. Table 7 below shows some lexical items as they are defined in the Bulgarian lexicon, in the Czech lexicon and in the Russian lexicon.

Table 7. Lexical items representation from the lexicons of Bulgarian, Czech and Russian resources respectively

```
(LEXICAL-ITEM
     :NAME       NACHERTAJA
     :SPELLING   "начертая"
     :FEATURES   (DO-VERB EFFECTIVE-VERB PERFECTIVE-VERB
                  DISPOSAL-VERB VERB))

(LEXICAL-ITEM
     :NAME       LINIA
     :SPELLING   "линия"
     :FEATURES   (NOUN COMMON-NOUN COUNTABLE FEM-NOUN))

(LEXICAL-ITEM
    :NAME       nakreslit
    :SPELLING   "nakreslit"
    :FEATURES (VERB PERFECTIVE DO-VERB EFFECTIVE-VERB
               DISPOSAL-VERB CREATION-VERB TRANSITIVE))

(LEXICAL-ITEM
    :NAME       u2sec3ka
    :SPELLING   "useeka"
    :FEATURES (NOUN COMMON-NOUN COUNTABLE FEMININE))

(LEXICAL-ITEM
     :NAME       NARISOVATJ
     :SPELLING   " narisovatj"
     :FEATURES   (PERFECT VERB DO-VERB EFFECTIVE-VERB
                  DISPOSAL-VERB)
     :PROPERTIES ((СВ 2 A)))

(LEXICAL-ITEM
     :NAME       LINIJA
     :SPELLING   " liniya "
     :FEATURES   (COUNTABLE COMMON-NOUN NOUN FEMIN)
     :PROPERTIES ((? 7 A)))
```

4.2. An example of multilingual generation in Bulgarian, Czech and Russian

The generation in AGILE system starts with conceptual specification of target text in so called A-box, which gives the structure and content of instructional procedure with

specified title, steps, side effects and so on. This conceptual specification is common for the three languages.

The next step in multilingual generation is the work of Text Structuring Module, which divides the text into clauses and prepares conceptual description for each clause. The semantic of each clause is given in the form of formal expression and that notation is (usually) common for the tree languages. The rules of building such expressions are developed by Kasper in 1989 as Sentence Plan Language.

The following example shows the Sentence Plan notation for the clause "Draw a line."

```
(S / DM::DRAW
:SPEECHACT   IMPERATIVE
:ACTOR   (HEARER / DM::USER)
:ACTEE  (AE / DM::LINE ))
```

Sentence Plan Language notation (SPL for short) contains Domain Model concepts making the meaning of the clause. SPL gives information about the particular concepts' roles in speech functions. Domain Model concepts in our particular example are DM::DRAW, DM::USER and DM::LINE. As it was explained above Domain Model definitions give the concepts particular meanings and map them to the Upper Model hierarchy. For example, DM::DRAW is defined as SIMPLE-ACTION (see Table 6), which is USER-ACTION-DIRECTED, which is DIRECTED-MATERIAL-ACTION, so that by nature DM::DRAW involves ACTOR and ACTEE roles (in respect to Upper Model definition of the later). Sentence Plan relates these two roles respectively to the Domain Model concepts DM::USER and DM::LINE, which have their particular meanings defined in Domain Model. SPEECHACT is also fixed in the Sentence Plan and the roles in this respect are given to the participants, too: : SPEECHACT is IMPERATIVE; the ACTOR of Domain Model concept DM::DRAW is associated to the HEARER of the IMPERATIVE SPEECHACT.

In general, the three views to the clause (clause as message, clause as representation and clause as exchange) are presented in the particular SPL.

In Fig. 1, Fig. 2 and Fig. 3 bellow are given the structures of the clauses automatically generated from this particular SPL notation in Bulgarian, Czech and Russian respectively.

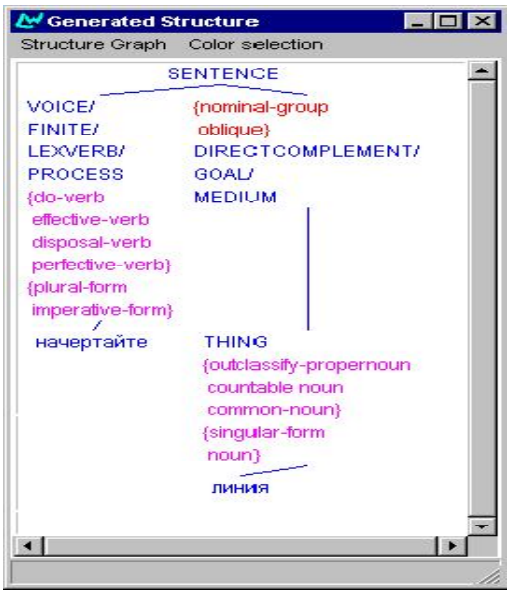

Fig. 1. Generated structure of the clause "Draw a line" in Bulgarian

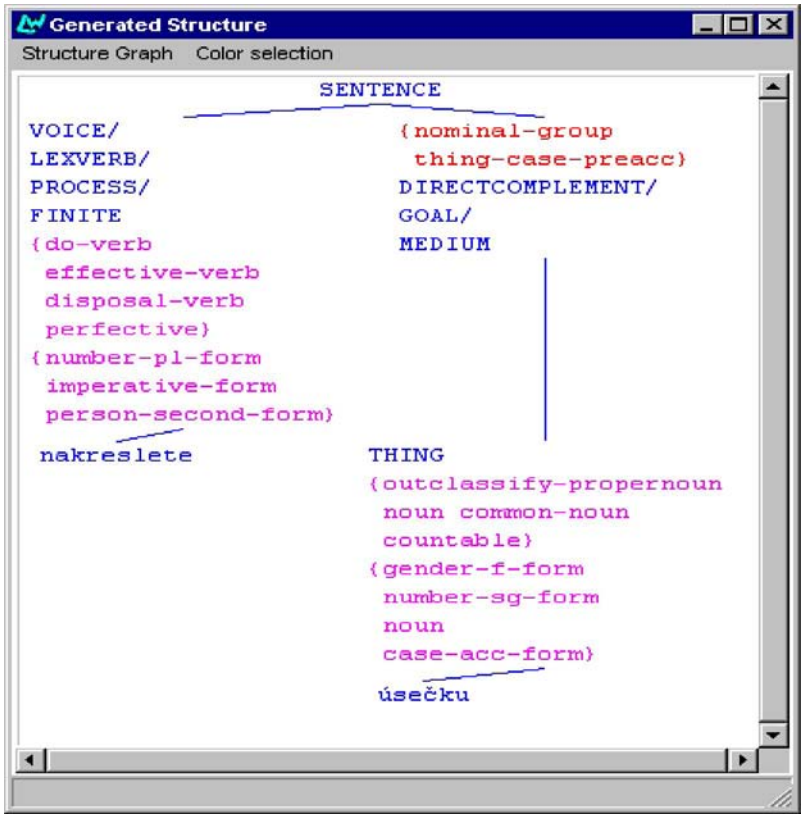

Fig. 2. Generated structure of the clause "Draw a line" in Czech

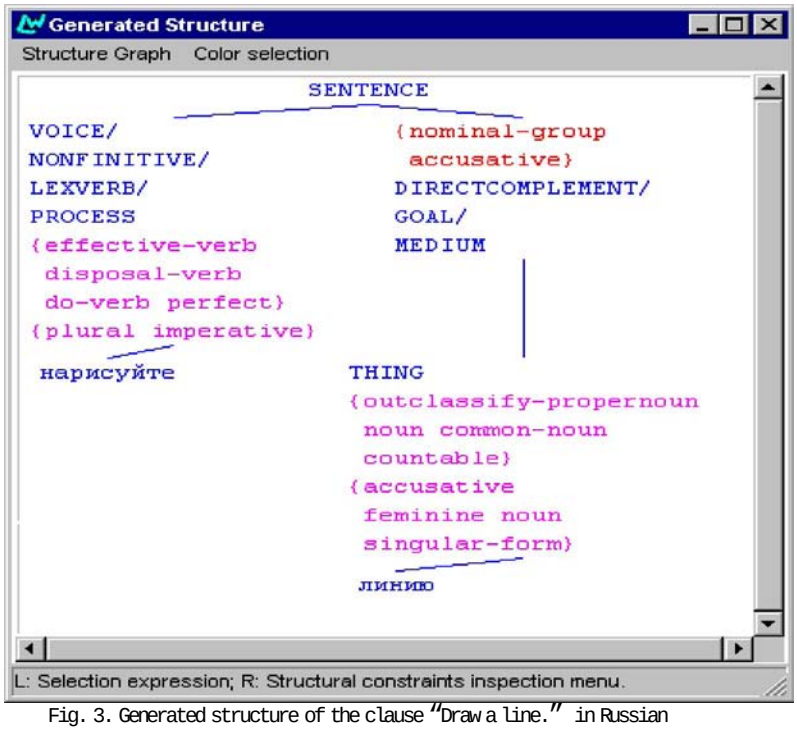

Fig. 3. Generated structure of the clause "Draw a line." in Russian

5 Future work

There are several directions of future work, which keep the author's attention and ambitions. The first one is the development of theoretical Systemic Functional Grammar of Bulgarian language and improving the existing Bulgarian computational resource to become "large-scaled" Bulgarian lexico-grammar. Interesting issue is the development of formal methodology for assessing Natural Language Generation resources and the further work on multilingual generation.

References

1. Halliday, M. A. K. An Introduction to Functional Grammar. Second Edition. London, Edward Arnold, 1994.
2. Martin, J. R., M. ICristian, M. Mattchissen, C. Painter. Working with Functional Grammar, St. Martin's Press, 1997.
3. Bateman, J.A., C. M. I. M. Matthiessen, K. Nanri, L. Zeng. The re-use of linguistic resources across languages in multilingual generation components. – In: Proc. of IJCAI'91, Sydney, Australia, **2**, Morgan Kaufmann Publishers, 966-971.
4. Bateman, J. A. Enabling technology for multilingual natural language generation: the KPML development environment. – In: Journal of Natural Language Engineering. **3**, 1997, No1, 15-55
5. Bateman, J. A., G-J. Teich, Iv. Kruijff-Korbayova, S. Sharoff, H. Skoumalova. Resource for multilingual text generation in three slavic languages. – In: Proc. of LREC' 2000, Athens, Greece.
6. Kruijff, G.-J., E. Teich, J. Bateman, I. Kruijff-Korbayova, H. Skoumalova, S. Sharoff, L. Sokolova, T. Hartley, K. Staykova, K., J. Hana. A multilingual system for text generation in three slavic languages. – In: Proceedings of the 18th Conference on Computational Linguistics (COLING 2000), Universitдt des Saarlandes, Saarbrucken, Germany, 2000, 474-480.
7. Rayner, M., D. Carter, P. Bouillon. Adapting the Core Language Engine to French and Spanish. – In: Proceedings of NLP-IA-96, Moncton, New Brunswick, 1996.
8. Bateman, J. A., R. T. Kasper, J. D. Moore, R. A. Whitney. A General Organization of Knowledge for Natural Language Processing: the Penman Upper Model. 1990.
9. Bateman, J. A., R. Henschel, F. Rinaldi. Generalized upper model 2.0: documentation. Technical report. GMD/Institut fuer Integrierte Publikations- und Informationssysteme. Darmstadt, Germany, 1995.
10. Bateman, J. A. The theoretical status of ontologies in natural language processing.– In: Proceedings of the workshop on 'Text Representation and Domain Modeling– Ideas from Linguistics and AI', Berlin, Germany, 1991.
11. Bateman, J. A., A. F. Hartley. Target suites for evaluating the coverage of texts generators. – In: Proceedings of LREC' 2000, Athens, Greece.
12. Bateman, J. A., S. Sharoff. Multilingual grammars and multilingual lexicons for multilingual text generation. – In: ECAI 98 13th Europian Conference on Artificial Intelligence, 1998.

Мультиязыковое генерирование фрагментов текстов (опыт из проэкта Agile)

Каменка Стойкова

Институт информационных технологий, 1113 София

(Р е з ю м е)

Представлены некоторые аспекты связанные с работой автора по теме международного проэкта для мультиязыкового генерирования фрагментов текстов. Концептуальная база разработки представляет формализмом представления лингвистических знаний, известной в области автоматического генерирования эстествено-языковых текстов (функциональная систематическая грамматика Халидея). Используется инструментальная система KPML и ресурсы, развитые в проэкте – онтологию, SFL лексико-грамматические ресурсы и речник с примерами для информационных структур. Показан пример генерирования простой клаузы на болгарском, чешком и русском языке.