БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ . BULGARIAN ACADEMY OF SCIENCES

ПРОБЛЕМИ НА ТЕХНИЧЕСКАТА КИБЕРНЕТИКА И РОБОТИКАТА, 49
PROBLEMS OF ENGINEERING CYBERNETICS AND ROBOTICS, 49

София . 2000 . Sofia

Lexico-Grammatical Characteristics of Bulgarian Software Instructional Texts[1]

Danail Dochev, Nevena Gromova**, Kamenka Staykova***

**Institute of Information Technologies, 1113 Sofia; New Bulgarian University, Sofia*

***Institute of Information Technologies, 1113 Sofia*

1. Introduction

The work described in this paper is performed under the international project AGILE whose aim is to develop a generic set of tools and linguistic resources for generating software instructional texts in CAD/CAM domain in Bulgarian, Czech and Russian. The project is based on the experience of the DRAFTER project, developed in the University of Brighton for English and French. One of the initial tasks of AGILE was to investigate the differences in the software instructional texts written in the three Slavic languages with respect to the languages analysed in DRAFTER. This was carried out by preparation of three parallel corpora of instructional texts for each of the Slavic languages, marked by a common tag set to reveal the correlation between the text unit functions and the chosen lexico-grammatical characteristics.

The corpus preparation and processing activities followed the methodology proposed in Hartley, Paris [3], consisting of the following steps:

1. Collection of texts from CAD/CAM software manuals.
2. Identification of linguistic features to be considered during the corpus analysis.
3. Marking of the corpus in terms of the selected features.
4. Computation of the frequency count of each linguistic feature.
5. Identification of the co-occurrences between linguistic features and text unit functions.

Corpus analysis was carried out within the framework of Systemic-Functional Linguistic (Halliday [2]), which views language as a resource for the creation of meaning.

[1] The research work described in the paper is funded under EC INCO-COPERRNICUS project PL961104 AGILE "Automatic generation of Instructions in Languages of Eastern Europe".

2. Corpus preparation and coding

2.1. Selection of corpus texts

Technical manuals within a specific domains constitute a sublanguage (Sager [8]). An important property of a sublanguage is its lexical and syntactical closure. It was shown (e.g. in Kitteridge [6]) that after the first 2000 words of a sublanguage text the number of new words increases little if at all. This lexical closure is determined by the domain specificity of the sublanguage, as well as by the norms of technical communication, which prefer monosemy to synonymy. The syntactic closure in sublanguages (leading to application of small number of rigid syntactical structures) was also demonstrated in several works.

Considering these results a small corpus volume was chosen for the needs of the project. A popular CAD/CAM manual (AutoCAD [1]) was selected as a source for the corpus preparation because of its wide use and availability. Nine texts from Chapter 2 of the manual describing procedures for producing graphical objects were chosen to be included in the parallel corpora. The selected texts are mostly procedural, corresponding to the specifications planned for the first two project phases - the Initial and Intermediate prototypes. As CAD/CAM manuals are a natural mixture of functional and procedural paragraphs some descriptive text chunks are also included in the corpora, though in limited number.

2.2. Purpose of coding and choice of the coding features

The corpus analysis was intended to reveal the necessary linguistic resources with respect to their use in software instructional texts. The corpus processing was used to help in the determination of:

- domain model concepts
- text planning process
- lexical resources
- grammatical resources

The choice of linguistic features for corpus tagging was based on the previous work (Hartley, Paris [3], Hartley, Paris [4]) done as part of the DRAFTER project. Its theoretical framework is the Systemic-Functional Linguistics (SFL) (Halliday [2]). The functional approach of SFL implies a high degree of commonality between the features of the analysed languages for the purposes of the AGILE project.

The corpus analysis was conducted on a set of software instructional texts which are characterised by a specific text structure and text plan. The text plan of the procedural text may be expressed by combination of the following elements: Goal, Precondition, Step, Side-effect, Interrupt. Each procedure has a goal and consists of a set of steps. Often in order to perform a procedure precondition(s) have to be satisfied. The normal stream of performed actions may be accompanied with side-effects/ intermediate results. The Interrupt element describes additional actions, necessary to undo the results of the current procedure. Each text to be generated has a communicative goal and intended meaning, which have to be realised by the most appropriate linguistic expression. In our case the concrete plan element is realised by some typical linguistic features.

The semantic function of the plan element allows a mapping from semantics to grammar. SFL is a grammar pushed to semantics, i.e. every grammatical feature encodes a specific meaning. The three functional components of meaning in SFL (metafunctions) are expressed by different grammatical structures. Each lexico-grammatical feature (LGF) is a language attribute with several possible values.

2.3. Tag set for AGILE corpora

The following list of attributes and their possible values were used for corpora coding:

- Text unit: Introduction, Procedure, Related-procedures, Further-possibilities, Other {**TI, TP, TRP, TFP, TO**}

These text units correspond to the specificity of the chosen procedural texts.

- Plan elements: Goal, Precondition, Step, Side-effect, Interrupt {**G, P, S, E, I**}
- Lexico-Grammatical Features
- **LGF 1**: Rank

{**CL** – Clause, **NGR** – Nominal Group, **PGR** – Prepositional group}

The RANK is a basic grammatical hierarchical unit in SFL. All of the metafunctions are combined in the clause, the unit with which SFL is concerned most.The semantic difference between the rank of the clause and the rank of the group is that in a group only one functional component of meaning is expressed. For instance, the experiential meaning of the ideational metafunction is expressed mainly in the nominal group, which is another feature in our tag set.

- **LGF 2**: Process-type

{**RP** – relational, **MNP** –mental, **VEP** – verbal,
DMP – directed-material, **NMP** – nondirected-material}

When a clause is viewed as representation, the function it expresses is ideational, i.e. modelling of experience. The basic components of such clause are process, participants and circumstances. Almost every clause represents some kind of process. The next six LGF are features closely related and dependent on the kind of process.

- **LGF 3**: Finiteness

{**FIN** – finite, **NFIN** – nonfinite}

- **LGF 4**: Polarity

{**POS** – positive, **NEG** – negative}

- **LGF 5**: Modality

{**NMD** – notmodal, **MDZP** – probability, **MDZU** – usuality,
MDLI – inclination, **MDLA** – ability, **MDLO** – obligation}

Some clauses have the meaning of exchange. The typical grammatical realisation of the interpersonal metafunction are the grammatical categories of Mood and Modality. The instructional texts as a whole have a clear interpersonal meaning, that of commands given to the user and request of action.

- **LGF 6**: Mood

{**IMP** – imperative, **INT** – interrogative, **IND** – indicative}

- **LGF 7**: Voice

{**VMID** – middle, **VEFA** – active, **VEFP** – passive}

- **LGF 8**: Agency-specified

{**AGSU** – agency-specified-user, **AGSP** – agency-specified-program,
AGSA – agency-specified-act, **AGN** – agency-notspecified, **AGG** – agency-specified-generic}

- **LGF 9**: Member-of-clause-complexity

{**CCS** – clause-complexity-simplex, **CCC** – clause-complexity-complex}

All the rest of the chosen lexico-grammatical features realise the textual metafunction: clause complexity – **LGF9** and clause interdependency – **LGF10**, clause taxis – **LGF11** and conjunction type – **LFG12**.

- **LGF 10**: Clause-dependency

{**NDC** – notdependent-clause, **DCT** – thematized-dependent-clause,
NDCT – notthematized-dependent-clause}

- **LGF 11**: Clause-taxis

{**PRC** – paratactic-relation clause, **HRC** – hypotactic-relation clause}

- **LGF 12**: Conjunction-type

{**CTAD** – conjunction-type-additive, **CTAL** – conjunction-type-alternative,
CTT – conjunction-type-temporal, **CTS** – conjunction-type-spatial,
CTM – conjunction-type-manner, **CTCR** – conjunction-type-causal-Reason, **CTCP** – conjunction-type-causal-purpose,
CTCC – conjunction-type-causal-condition}

2.4. Description and content

Each corpus consists of two parts:

1. Text file containing the sequentially numbered coding units.
2. Coding table containing the feature values for each coding unit.

As Bulgarian translation of the original AutoCAD User's Guide (AutoCAD [1]) is not available, the corpus texts were translated especially for the purposes of the AGILE project. A Bulgarian translation of another AutoCAD manual (Zirbel, Combs [9]) was consulted during the translation. The corpus contains nine procedural texts with 1025 words and 194 coding units (Hartley, Dochev et al. [5]).

The results of corpus analysis are summarised in two tables of co-occurrences - one table for the whole corpus and one table (a subset of the first one), covering only the procedure elements of the text plan (Goal, Step and Side-effect). This structuring of analysis results was chosen to highlight more clearly the intended differences in the strictly procedural chunks of text and the more descriptive functional chunks of text, expressing Precondition and Interrupt plan elements.

2.5. Tabular summary of co-occurrences

There are given two tables of co-occurrences for the whole corpus and for the procedures only.

2.5.1. Table of the co-occurrences for the whole corpus

	G	P	S	E	I
Rank					
CL	83.02%	50%	100%	100%	100%
NGR	16.98%	50%	0%	0%	0%
PGR	0%	0%	0%	0%	0%
Process-type					
RP	0%	7.69%	1.03%	0%	0%
MNP	1.89%	7.69%	0%	0%	0%
VEP	0%	0%	0%	0%	0%
DMP	92.45%	84.62%	98.97%	70%	100%
NMP	5.66%	0%	0%	30%	0%
Finiteness					
FIN	100%	100%	96.88%	100%	100%
NFIN	0%	0%	3.12%	0%	0%
Polarity					
POS	100%	92.86%	100%	98.97%	100%
NEG	0%	7.14%	0%	1.03%	0%

	G	P	S	E	I
Modality					
NMD	90.91%	85.72%	97.92%	80%	100%
MDZP	0%	0%	0%	0%	0%
MDZU	0%	0%	0%	0%	0%
MDLI	2.27%	7.14%	0%	0%	0%
MDLA	6.82%	7.14%	1.04%	20%	0%
MDLO	0%	0%	1.04%	0%	0%
Mood					
IMP	13.64%	0%	86.46%	0%	60%
INT	0%	0%	0%	0%	0%
IND	86.36%	100%	13.54%	100%	40%
Voice					
VMID	2.27%	0%	0%	20%	0%
VEFA	97.73%	100%	100%	80%	100%
VEFP	0%	0%	0%	0%	0%
Agency					
AGSU	93.18%	85.72%	98.96%	50%	100%
AGSP	6.72%	4.28%	1.04%	50%	0%
AGSA	0%	0%	0%	0%	0%
AGN	0%	0%	0%	0%	0%
AGG	0%	0%	0%	0%	0%
Clause compl.					
CCS	2.27%	13.33%	42.71%	20%	0%
CCC	97.73%	86.67%	57.29%	80%	100%
Clause depend.					
NDC	22.73%	20%	88.54%	80%	60%
DCT	15.91%	46.67%	0%	0%	20%
NDCT	61.36%	33.33%	11.46%	20%	20%
Clause taxis					
PRC	8.82%	8.33%	52.17%	50%	33.33%
HRC	91.18%	91.67%	47.83%	50%	66.67%
Conjunction					
CTAD	5.88%	0%	30%	50%	0%
CTAL	2.95%	10%	30%	0%	33.33%

2.5.2. Table of co-occurrences for the whole corpus

CTT	0%	60%	0%	0%	0%
CTS	0%	0%	5%	0%	0%
CTM	0%	0%	35%	0%	0%
CTCR	0%	0%	0%	0%	0%
CTCP	85.29%	0%	0%	50%	66.67%
CTCC	5.88%	30%	0%	0%	0

Table of co-occurrences for procedures only

	G	S	E
Rank			
CL	83.02%	100%	100%
NGR	16.98%	0%	0%
PGR	0%	0%	0%
Process-type			
RP	0%	1.03%	0%
MNP	1.89%	0%	0%
VEP	0%	0%	0%
DMP	92.45%	98.97%	70%
NMP	5.66%	0%	30%
Finiteness			
FIN	100%	96.88%	100%
NFIN	0%	3.12%	0%
Polarity			
POS	100%	100%	98.97%
NEG	0%	0%	1.03%
Modality			
NMD	90.91%	97.92%	80%
MDZP	0%	0%	0%
MDZU	0%	0%	0%
MDLI	2.27%	0%	0%
MDLA	6.82%	1.04%	20%
MDLO	0%	1.04%	0%
Mood			
IMP	13.64%	86.46%	0%
INT	0%	0%	0%
IND	86.36%	13.54%	100%
Voice			
VMID	2.27%	0%	20%
VEFA	97.73%	100%	80%
VEFP	0%	0%	0%
Agency			
AGSU	93.18%	98.96%	50%
AGSP	6.72%	1.04%	50%
AGSA	0%	0%	0%
AGN	0%	0%	0%
AGG	0%	0%	0%
Clause compl.			
CCS	2.27%	42.71%	20%
CCC	97.73%	57.29%	80%
Clause depend.			
NDC	22.73%	88.54%	80%
DCT	15.91%	0%	0%
NDCT	61.36%	11.46%	20%
Clause taxis			
PRC	8.82%	52.17%	50%
HRC	91.18%	47.83%	50%
Conjunction			
CTAD	5.88%	30%	50%
CTAL	2.95%	30%	0%
CTT	0%	0%	0%
CTS	0%	5%	0%
CTM	0%	35%	0%
CTCR	0%	0%	0%
CTCP	85.29%	0%	50%
CTCC	5.88%	0%	0%

The analysis of co-occurrence tables allows making the following conclusions about the sublanguage used in Bulgarian software manuals:

- The great majority of the rank units are clauses and the rest are nominal groups. The prepositional groups do not occur in instructional texts of the corpus.
- The processes are exclusively of the directed-material type. Sometimes relational, mental, and not-material processes are found in the corpus. The only kind of process not present in it is verbal process.
- Finite and positive polarity predominate over non-finite and negative polarity features in this particular sublanguage.
- Most of the analysed clauses are non-modal. In the case when modality is expressed in the clause it is of the ability, inclination and obligation type.
- The mood is usually realised by an imperative clause.
- The voice is active, although a few instances of middle were counted. The passive voice was not found in instructional texts at all.
- The user is the most frequent agent, the alternative is program objects appearing as agents.
- Most of the text units are members of a complex clause. The interdependency between the complex clauses is as likely to be paratactic as hypotactic. The hypotactic relation is realised mainly by a manner or purpose conjunction, although condition and temporal conjunctions occur as well. Paratactic relation is realised by the additive and alternative conjunctions.

The summary table of the occurrence of a particular lexico-grammatical feature in each plan element shows the following correlation between plan elements and lexico-grammatical features:

- The Goal is realized mostly by a clause (83%), the alternative realization is by a nominal group (17%) particularly for top-level goals. In the Precondition plan element the distribution between clause (50%) and nominal group (50%) is even. Steps, Effect and Interrupt are realized exclusively be clauses (100%).
- Directed-material process overwhelmingly expresses all the plan elements. It is the only type of process for Interrupt (100%). There are few occurrences of mental process in Goal (2%) and Precondition (8%), and non-material process in Goal (6%) and Effect (30%).
- The plan elements are expressed usually by means of a finite clause, just 3,2% of Step are realized by a nonfinite clause. The figures are similar for Polarity, since the plan elements are realized mainly by positive clauses, with the exception of Precondition, which admits negative polarity in 7%, and Effect in 1% of the cases.
- The overwhelming majority of the clauses in our corpus have no expression for modality. There are a few instances of inclination in Goal (2%) and Precondition (7%) and ability again in Goal (7%) and Precondition (7%). Ability hardly occurs in Step (1%), but is quite frequent in Effect (20%). Interrupt is the only plan element which does not admit modals at all (100% nonmodal).
- The mood in Goal (86%), Precondition (100%), Effect (100%) is imperative, and in Step (86%) and Interrupt (60%) is indicative. There are no occurrences of interrogative mood in procedural texts.
- The most frequent expression for the plan elements in terms of voice is active voice, the alternative for active voice in Goal and Effect is middle voice. Instances of

passive voice were not found in procedures.

- The user is the agent of the clause in 93% in Goal, 86% in Precondition, 99% in Step, and 100% in Interrupt. The agent can be either user (50%) or program (50%) in Effect.
- All the clauses in Interrupt are complex clauses (100%). In the rest of the plan elements both kinds of clauses appear. The Goal is expressed by a simple clause in 2% of the cases, the Precondition in 13%, the Step in 42% and the Effect in 20%.
- Most of the clauses in Goal are notthematized dependent clauses (61%). In Precondition thematized dependent clauses predominate (47%). In Step there are no occurrences of thematized dependent clauses, mainly notdependent clauses (89%) or notthematized dependent clauses (11%) express the interdependency between clauses. Effect is never realized by thematized dependent clauses, most of the instances are notdependent clauses (80%) and 20% are realized by nothematized dependent clauses. In Interrupt the correlation is notdependent clauses (60%) versus thematized dependent (20%) and notthematized dependent clauses (20%).
- Goals are usually expressed by paratactic additive (6%) or alternative (3%) conjunctions, as well as by hypotactic purpose (85%) or condition (6%) conjunction. Precondition is realised by alternative (10%), temporal (60%), condition (30%) conjunctions. The usual conjunctions appearing in Step are additive (30%), alternative (30%), spatial (5%) or manner (35%). The realization of Effect in terms of conjunction is distributed evenly between additive paratactic (50%) and purpose hypotactic (50%). Most conjunctions in Interrupt are hypotactic expressing purpose (66%), some of the conjunctions are alternative (33%).

The comparison of 2.5.1 and 2.5.2 shows the following differences between the LGF distribution over strictly procedural plan elements procedure-related plan elements (Precondition, Interrupt):

- While the text units in the corpus are mostly clauses, in the Precondition plan element the distribution between clause (50%) and nominal group (50%) is even.
- In contrast to the other plan elements Interrupt is the only one which does not admit modals at all (100% nonmodal).
- In contrast to the other plan elements Interrupt text chunks use indicative mood (60%).
- All the clauses in Interrupt text chunks are complex clauses (100%).

References

1. AutoCAD. AutoCAD Release 13. User’s guide, chapter 2. Autodesk Co., 1996.
2. Halliday, M.A.K. An Introduction to Functional Grammar. Second edition. London, Arnold, 1994.
3. Hartley, A., C. Paris. Two Sources of Control over the Generation of Software Instructions. – Information Technology Research Institute Technical Report Series, ITRI-96-02, 1996.
4. Hartley, A., C. Paris. Towards an Empirical Classification of the Discourse of Software Instructions. In: Computational Linguistic, **24**, 1998.
5. Hartley, A., D. Dochev, N. Gromova, K. Staykova, A. Bemova, A. Rosen, J. Trojanek, E. Sokolova. Tagging and analysis of instructional texts in the software domain. AGILE WP3 Deliverable, 1998.
6. Kitteridge, R. The Significance of Sublanguage for Automatic Translation. – In: S. Nirenburg (ed.). Machine Translation: Theoretical and Methodological Issues,1987.
7. Mitkov, R. The Sublanguage Approach: A Key to Realistic Natural Language Processing. – In: Proc. of. AIMSA’94, World Scientific, 1994, 391-400.

8. Sager, J. S. Language Engineering and Translation: Consequences of Automation. Wiesbaden, Germany, Brandstetter Verlag, 1993.

9. Zirbel, J.H., S.B. Combs. AutoCAD Release 13 for Windows. Professional edition, Part 1. Bulgarian language edition. SoftPress Publishing, 1996.

Лексико-грамматические характеристики руководств для программных продуктов на болгарском языке

Данаил Дочев, Невена Громова, Каменка Стойкова

Институт информационных технологий, 1113 София

(Резюме)

Исследуются лексико-грамматические характеристики руководств на болгарском языке. Используется примерный текст из руководства CAD-CAM системы. При помощи формализма представления лингвистических знаний, известного в области автоматического генерирования текстов как функциональная систематическая граматика Халидей, показаны 12 характеристик в таблицах. Сделан подробный анализ полученных результатов.