

LEXICAL-FUNCTIONAL GRAMMAR OF THE CROATIAN LANGUAGE: THEORETICAL AND PRACTICAL MODELS

SANJA SELJAN, Ph.D.

FACULTY OF PHILOSOPHY, DEPARTMENT OF INFORMATION SCIENCES
IVANA LUČIĆA 3, 10 000 ZAGREB, CROATIA, E-MAIL: SSELJAN@FFZG.HR

Formal description aims to find the most suitable way to formalize a certain segment of the language, or some language phenomena at morphological, lexical, syntactic or semantic level.

There is a variety of formal models based on different approaches in order to describe the natural language as much as possible. One type of such formal model is described in the thesis entitled "Lexical-Functional Grammar of the Croatian Language: Theoretical and Practical Models".

The aim of this work was:

a) To find the theoretical formal models and to define formal rules in order to describe certain language phenomena at the morphological, lexical, syntactic and semantic level for the subset of Croatian language sentences.

b) To verify computationally theoretical models using LFGW program in Amzi Prolog

The method used for the formal description of the Croatian is Lexical-Functional Grammar which is generative non-transformational grammar, belonging to the group of Unification Grammars.

The hypothesis to verify is definition of the LFG model as context-sensitive grammar or the grammar on the basis of constraints according to which it should be suitable for introduction of new characteristic features and for the description of language with relatively free word order.

1 Introduction

The formal model of the Lexical-Functional Grammar is used for the formal description of some segments of the Croatian language. Since the LFG model tends to belong to the context-sensitive type of grammars enabling agreement and introduction of new characteristic features, it is supposed to be suitable for description of the language with rich morphological system and relatively free word order, such as Croatian.

Because of the rich morphological system, case marking and agreement are one of the most important tasks to accomplish. Because of formal problems, some new parts of speech are introduced, as well as new characteristic features. Other linguistic phenomena are also analyzed, such as subcategorization, active and passive sentences, idioms, impersonal constructions, composite tenses, extraction and long-distance dependencies and clitics.

Since the LFG model tends to unify computer efficacy with linguistic theories, and uses some methods from mathematics and logic, this model has its computer application. LFG has been used in the automatic text processing, in text generation and in machine translation.

The practical part of the LFG model consists of the computer program written in LFGW (LFG for Windows, code in Amzi Prolog) by A. Andrews, University of Brisbane. The program gives for the sentence in the Croatian two structures:

a) Constituent structure using tree form and defining the syntactic level with parts of speech and cases

b) Functional structure in the matrix form unifying information from the constituent structure and from the lexicon, which has to satisfy principles of uniqueness, coherence and completeness.

Sentences are firstly generated by syntactic rules, passing then constraining tests introduced in the lexicon or added in generative rules. The program does not represent the best solution (especially on the morphological level), but sentences are analyzed on all linguistic levels.

2 Why LFG model?

The formal model should be suitable not only for highly configurational type of languages, but also for nonconfigurational languages with relatively free word order providing simple morphological analyzer (J.BRESNAN, 2001: 6-10, "Morphology competes with syntax."). Therefore, besides the representation suitable for the specific natural languages (principle of variability), a more abstract representation is also needed (principle of universality). This means that the formal grammar should satisfy several conditions: to be context-sensitive, and to be suitable for highly structured as well as for nonconfigurational types of languages.

Having in mind that the perfect formal model doesn't exist (because of ambiguity, complexity and unlimited possibilities of combinations in the language) and that formal descriptions are used for a certain subset of the language sentences or for a controlled language, there are several reasons for the LFG model to be chosen.

Besides problems valid for all natural languages, the Croatian language is characterized by relatively free word order conditioned by rich morphological system. Since the words in the Croatian are quite strongly connected among themselves (case marking and agreement being ones of the most important elements), it is necessary to use context-sensitive grammar (although the perfect one still does not exist) which could analyze words in the environment of the left and right context, and thus perform not only syntactic, but also semantic analysis (although in a quite restricted way).

According to the principle of the Universal Grammar, the basic principle of the LFG model is to use grammatical functions presented in lexicon. Assuming that despite very different syntactic means of expression all natural languages have grammatical functions as subject and object, the central role belongs to the grammatical functions and not syntactic categories. The idea enables categorial independency, i.e. that different word categories can have the same grammatical function and than one category can have different grammatical functions.

Lexical-Functional Grammar is generative non-transformational grammar. Belonging to the group of Unification Grammars, that use the operation of unification as a principal operation, it enables unification of characteristic features.

LFG formal model uses parallelly two basic structures: constituent structure describing constituent parts, i.e. the surface structure which is language specific and functional which is more abstract, and tends to be universal for the same sentence in different languages (although other structure have been added, such as argument and morphological structures).

Decomposition of categories on characteristic features enables incorporation of contextual elements and agreement, which is in meta-language reflected as attribute-value pairs. This way, it is possible to introduce new language specific characteristic features, necessary for description of some language phenomena (for instrumental case $Thg=+$ and $Soc=+$, for agreement of collective nouns $Coll=+$, for reflexive verbs $Refl=+$; demonstratives have formal feature of three levels of proximity $PROX= 1/ 2/ 3$, i.e. *ovaj, taj, onaj* – in English *this, that*).

Since LFG model has been developed by linguists and information specialists, aiming to unify linguistic theory and computer efficiency (J. Bresnan & R. Kaplan), it has been applied for formal description of different language phenomena in various types of languages (English, French, German, Italian, Spanish, Irish, Dutch, Russian, Arabian, Japanese, Urdu, Walpiri, Bantu languages, Icelandic, Flemish, Norwegian, Malayalam, Moroccan, etc.) and has its computer implementation.

3 Croatian language

Besides general problems, every language has its own requirements and, correspondingly, its specific problems regarding its description.

The Croatian language, as all Slavic languages, belongs to the group of highly fleective languages with rich morphological system. One word may have a great number of varieties (taking for example the root *prijatelj* (eng. *friend*), from which a considerable number of different words can be derived: singular and plural in 7 cases of the word *prijatelj* (eng. *friend*), singular and plural in 7 cases of the word *prijateljica* (eng. *woman friend*), singular and plural in 7 cases of the adjective *prijateljski*, -a, -o in 3 genders (eng. *friendly*), adverb *prijateljski* (eng. *friendly way*), common nouns such as *prijateljstvo* (eng. *friendship*), *prijateljstvo* (eng. *friendly relations*) in singular and plural in 7 cases, verb *prijateljivati* (eng. *to be on friendly terms with someone*), etc.

Suffixes and endings denoting the parentage, gender, number, case, person etc. are simply added on the root of the word (although there are differences in distinguishing what the root is). According to some researches, 100.000 varieties in the English language correspond to approximately 1.000.000 (one million) varieties in the Croatian language.

In the Croatian language endings are strongly connected to the root of the word and to the suffix, conditioning thus relatively free word order. Although the most common and non-marked in the style is SVO (Subject-Verb-Object) construction, other constructions are possible as well, but these are then marked in the style. In the SVO construction, all syntactic categories are equally important, that is, none is specially stressed, while in the stylistically marked word-order, all word groups are not equally important. The ones being at the beginning or at the end of the sentence are stressed.

The sentence which in English or French language may be expressed by only one type of word-order (*I am writing a book*, fr. *J'écris le livre*) has several possibilities in the Croatian (*Ja pišem knjigu*, *Pišem knjigu* *Knjigu pišem*, etc.), the most neutral being the one with SVO construction. In the mentioned sentence the person is not to be mentioned necessarily unless it is stressed by the ending *-em* which denotes the person in question (first person, singular). Thus variations regarding the word order are possible, although the SVO construction is the most common and neutral.

3.1 Morphology competes with syntax

Fleective languages, like Croatian, compensate highly structural order by rich morphological system, that enable changes of syntactic groups, which is marked by flat structure, although hierarchical structure is also possible, marking preferred, stylistically neutral structure.

The flat structure does not have elements such as specifier – head – complement where X'-theory is not applied. According to Y. FALK. (2001:50), constituent structure without head is called *exocentric*, and the language generated by this structure are called W* languages, because of the rule S→W* where W marks any type of word.

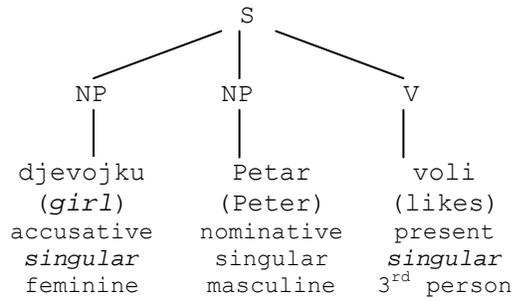


Figure 1. Flat constituent structure for Croatian sentence

J. BRESNAN (2001:98) talks about *endocentric and lexocentric* constituent structure; endocentric organization appearing in highly structured tree composition and lexocentric in flat constituent structure, where all arguments appear as sister nodes of the verb and syntactic functions are defined by morphology, i.e. by cases and agreement. Endocentric organization is defined by X'-theory, where X is a head of X', X' head of X". In lexocentric organization, grammatical functions are encoded by lexical means, such as case and agreement morphology. Most of languages have a mixture of endocentric and lexocentric organization.

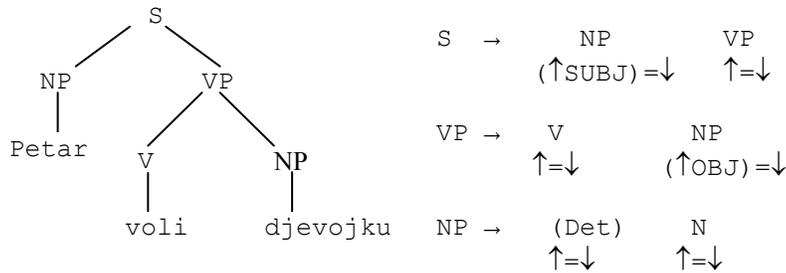


Figure 2. Hierarchical constituent structure for Croatian sentence

According to LFG model, different constituent structures can have one common structure, *functional structure*, which is more abstract and represent the meaning of the sentence.

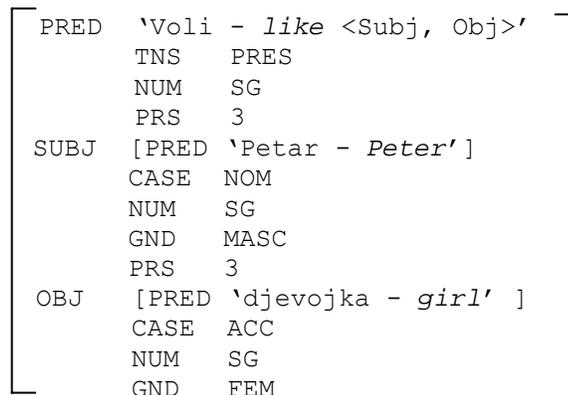


Figure 3. Functional structure

M. TALLERMAN (1998:146-150) numerates six basic types of word order, inside o which three constituents are distinguished: subject (S), verb (V), object (O): SVO (*Ivica voli Maricu – eng. John – loves - Mary*), SOV (*Ivica Maricu voli*), VSO (*Voli Ivica Maricu*), VOS (*Voli Maricu Ivica*), OVS (*Maricu voli Ivica*), OSV (*Maricu Ivica voli*).

M. Tallerman indicates that the first two examples SVO and SOV and stylishly unmarked and used in 80-90% languages in the world. VSO is the second one according to its extent (9-12%), while VOS is used in only 3% of languages. In most languages, subject is in the initial position and in 96% subjects precedes object. In 90% verb is next to the object (VO or OV construction).

4 Levels of representation

According to the traditional LFG literature, this formal model supports two basic level of representation (*constituent and functional*), although the others are closely related and afterwards distinguished as separate levels of representation. In 1989, Bresnan and Kanerva added *argument structure* as a transitive structure between constituent and functional structure. *Morphological level* has been topic of discussions in the last several years. There are five structures that can be distinguished, although not strictly separated. One of the reasons for choosing LFG model is that uses several levels of representation simultaneously:

(1) Since grammatical functions are represented in the lexicon, *lexical level* is of the crucial importance, including information about meaning of the item, its argument structure and grammatical functions. Grammatical functions (such as subject, object, etc.) play essential role and mediate between lexicon and syntax.

(2) *Constituent structure* reflecting the *syntactic structure*, encoding linear order, hierarchy and syntactic categories of constituents (in the form of *context-free ruled enriched with functional annotations* or in the form of the tree). Constituent structure varies from one language to another.

C-structure corresponds to the superficial phrase structure and works closely with an enriched lexical component. C-structure exists simultaneously with f-structure that integrates information from the lexicon and c-structure.

(3) *Argument structure* has been added afterward as a separate structure (before 1989. it was included in the lexical structure), consisting of predicate and its arguments. According to Function-Argument Biuniqueness, every grammatical function (subject, object) can have one thematic role (agent, theme, goal, etc.)

A-structure was contained in the lexical level, but afterwards it has been separated as a transitive structure between c- and f- structures, pointing out the assignment of grammatical functions (e.g. subject, object, complements) to thematic roles (agent, theme, goal, etc).

(4) *Functional structure* which is more abstract and tends to be *universal* for the same sentence across the languages. It is represented in the form of matrix, integrating structural and lexical information, where *grammatical functions* are included, *regardless the position in the sentence*.

(5) *Morphological structure* that contains information about morphological form of auxiliaries enabling flat functional structure.

4.1 Lexical level

Lexicon is in LFG model the central point containing grammatical relations between predicate-argument structure and grammatical functions. Lexicon consists of lexical entries related to the paradigms.

The lexical entry includes different type of information:

- form of the item
- syntactic category (N, V, Adj, etc.)
- functional schemata containing information about meaning inside of quotes ‘ ‘ and grammatical functions (*subject, object etc.*) interrelated with thematic roles (*agent, theme, etc.*)

voli V (↑PRED)= 'voljeti - like <(↑SUBJ) (↑OBJ)>
 (↓NUM) = SG
 (↓PRS) = 3
 (↓TNS) = PRES

4.2 Constituent (C-) structure

C-structure encodes the linear order, hierarchy and syntactic categories. This structure is specific for every language.

C-structure can be presented in two ways:

- (1) by context-free rules that enriched by annotations
- (2) by the annotated phrase tree structure satisfying relations of precedence and domination

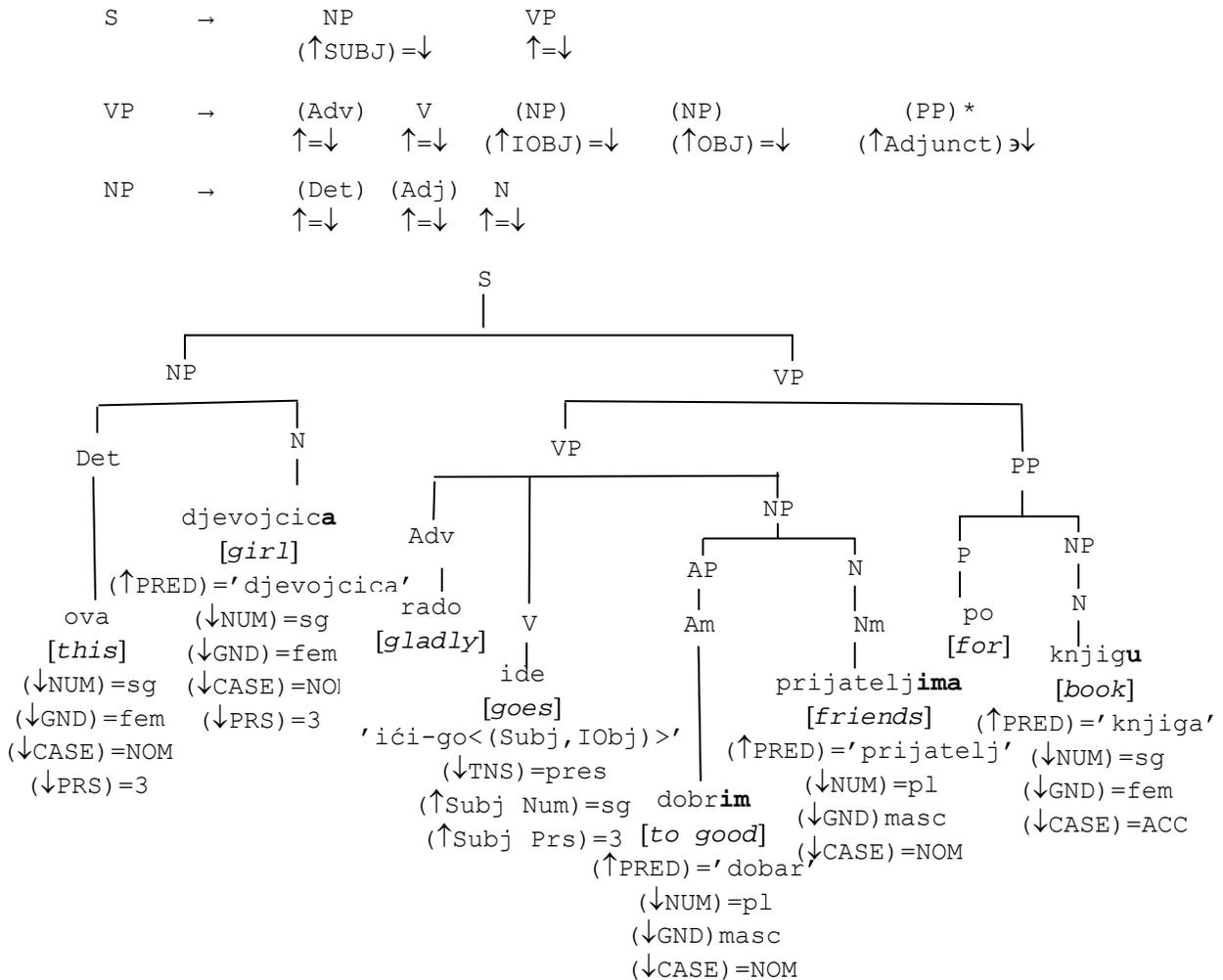


Figure 4. Annotated constituent structure

4.3 Argument (A-) structure

One of the basic assumptions of LFG is that grammatical functions are regulated through the predicate-argument structure found in the semantic form paired with PRED. Semantic forms appear graphically as material flanked with single quotes as PRED='djevojčica - girl', PRED='ići - go (↑SUBJ)(↑IOBJ)'. Verbs always incorporate predicate-argument structure.

Argument structure has been in 1989. by Bresnan and Kanerva as a separate transitive, structure between c- and f-structures. According to the *Functional-Argument Biuniqueness* principle each grammatical function (*Subj, Obj, IObj, etc.*) can be associated with only one thematic role (*Agent, Theme, Goal, etc.*). If we the sentence is transformed from active voice into passive, the principle of must be valid.

<i>Tom voli Anu.</i>		(SUBJ) (OBJ)
<i>Tom likes Ana .</i>	'voljeti-like	(agent, theme)'
<i>Marko je voljen.</i>		(Ø) (SUBJ)
<i>Marco is loved.</i>	voljen-liked	(agent, theme)'

The verb '*voljeti*' is a two-place predicate, where the PRED feature has as its value the meaning of the verb, subcategorizing subject and object related to roles of Agent and Theme.

In the case that agent is not alive, i.e. non-thematic or that agent is not directly related to the subject, it is placed outside the signs <>.

<i>On ju smatra ozbiljnim kandidatom.</i>		
<i>He is considering her serious candidate.</i>		
smatrati V	(↑PRED) = 'smatrati-consider	< (↑SUBJ) (↑XCOMP) > (OBJ)'
	(↑OBJ) = (↑XCOMP SUBJ)	

4.4 Functional (F-) structure

Kaplan suggests the functional structure for more abstract representation and considers grammatical functions independently from the position of words in the sentence, which is especially suitable for languages with free word order.

F-structure integrates lexical information from the lexicon and structural information from the c-structure, in the way that the lexical item placed as a terminal node of the tree characterized by information marked in the lexicon.

F-structure is presented in the form of hierarchically organized attribute-value matrix. Each pair of attribute and its value creates one characteristic feature.

Attributes may have three kinds of values:

- atomic symbols as [NUM PL] [PRS 3]
- semantic form which is indicated as the value of PRED and enclosed within '...'
PRED '*pokloniti-give*<(↑SUBJ)(↑OBL_{DAT})(↑OBJ)>'
- one or more subsidiary f-structures: a value of the attributes SUBJ and OBJ is again new f-structures consisting of attributes and values. The value of that attribute can be f-structure again, composed of attributes and atomic symbols. This is the case with subordinate clauses inside of which there is again new subordinate clause.

To be valid, f-structure must satisfy tree well-formed conditions: completeness, coherence and consistency principles.

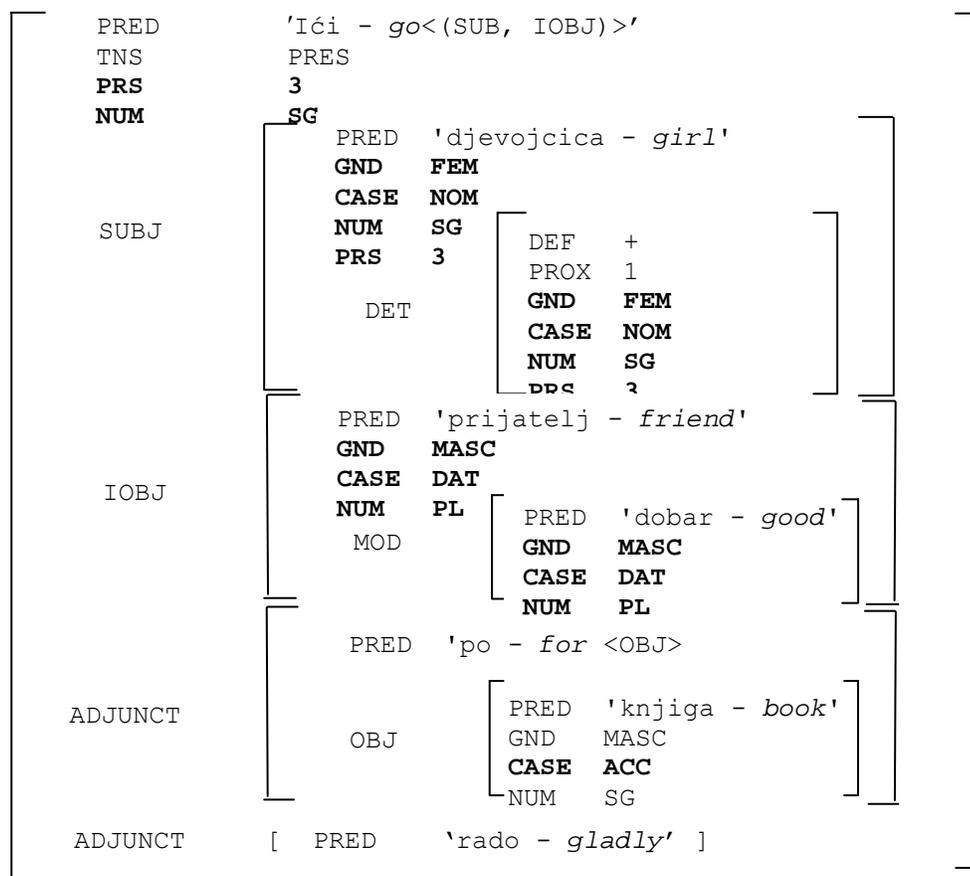


Figure 5. Functional structure

4.5 Morphological (M-) structure

Morphological structure is subject of many recent discussions (J. BRESNAN, 2001; A. FRANK, 2000; R. KAPLAN, 2002; M. BUTT, 2001) suggesting that information about auxiliaries, as well as information about strong forms (Str=+) of agreement features, should be contained in the m-structure. The m-structure would not contain PRED features and subcategorization frame with XCOMP function.

Therefore the same sentence on different languages (*Eng. She will read the book, Germ. Sie wird das Buch lesen. Fr. Elle lira le livre. Cro. Ona će čitati knjigu*) would have the same functional structure parallelly with morphological structure carrying information about auxiliary verb and of the related basic form (*will read, wird lesen, će čitati*). This model of presentation would enable description of cases when the main verb precedes the auxiliary what is in Croatian case (*Čitat će knjigu.*).

In that way morphosyntactic characteristics should not be introduced in the functional structure, which could be unique regardless synthetic or analytic forms (which would be presented in the constituent structure) in tense presenting. The f-structure would contain TNS-APS features.

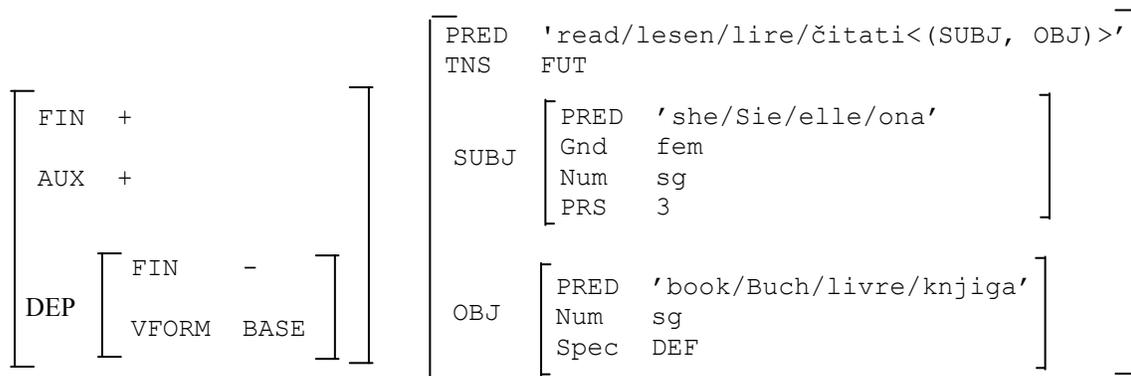


Figure 6. Morphological and Functional structures

5 Conclusion

Lexical-Functional Grammar is generative non-transformational grammar, formally characterized as context-sensitive grammar. Belonging to the group of Unification Grammars, that use the operation of unification as the principal one, it enables unification of characteristic features. Being context sensitive grammar based on the principle of grammatical functions presented in the lexicon, this formal grammar aims to be suitable for the description of various language phenomena in different types of languages. Decomposition on characteristic features is possible, as well as introduction of new features and contextual elements, which are in meta-language reflected as attribute-value pairs.

LFG model uses two main levels of representation: the constitutional c-structure for syntactic representation that is closely related to the lexicon, and the functional f-structure suitable for more abstract representation, where information on structural, syntactic and semantic level are merged. C-structure is represented in the form of tree with annotated rules, while f-structure is represented in the matrix form composed of attribute-value pairs. While c-structure varies from one language to another, f-structure tends to be universal. Another levels of representation are also presented, such as argument and morphological levels.

Since the aim of this dissertation is to find possible solutions (surely not the best, but possible ones) using LFG formal model, sentences are analyzed on morphological, lexical, syntactic and semantic level. Each of the level is described on the Croatian language and on English or French, pointing out analogy or contrast in the formal analysis.

In the formalization process, some new types of words, subgroups and characteristic features are introduced. Most of theoretical models have practical realization using LFGW program. The following language phenomena of the Croatian language have been described: case-marking, agreement, subcategorization, controlling principles for infinitive constructions, preterit and future, certain interrogative and relative sentences, negation, some coordination and passive sentences etc.

Although the formal LFG model does not offer solutions for complete analysis of the natural language sentences (e.g. coordination structures, ambiguities, some long-distance dependencies etc.), it does represent a useful step forward in the field of Natural Language Processing. As one of possible models representing attempt for formal description of the Croatian language phenomena, it can be viewed as the bridge between informatics, linguistics, logic and mathematics helping us to better understand the proper language in order to approach the theoretical models and practical computer application.

References:

- ABEILLÉ, ANNE. *Les nouvelles syntaxes: Grammaires d'unification et analyse du Français*. Paris: Armand Colin, 1993.
- ANDREWS, AVERY. *LFGW System*. University of Brisbane. <http://www-csli.stanford.edu/~andrews/lfgw.html>
- AUSTIN, P. Lexical-Functional Grammar. // *International Encyclopedia of the Social and Behavioural Sciences*. Smelser, Neil J.; Baltes, Paul, eds. Elsevier, 2001. p. 8748 – 8754. (<http://www.linguistics.unimelb.edu.au/contact/staff/peter/Elsevier.pdf>)
- BUTT, MIRIAM. The Treatment of Tense. // *Proceedings of the LFG01 Conference*. CSLI Online Publications. <http://csli-publications.stanford.edu/LFG/6/lfg01butt.pdf>
- BUTT, MIRIAM; DIPPER, STEPHANIE; FRANK, ANETTE; HOLLOWAY KING, TRACY. Writing Large-Scale Parallel Grammars for English, French and German. // *Proceedings of the LFG99 Conference*. <ftp://ftp.ims.uni-stuttgart.de/pub/users/dipper/papers/lfg99.pdf>
- CHOMSKY, NOAM. *Aspects of the Theory of Syntax*. Cambridge: MIT Press, 1965.
- CHOMSKY, NOAM. *Syntactic Structures*. Paris: Mouton, 1972.
- CLÉMENT, LIONEL. *Interactive demo with a Toy French Grammar*. Xlfg version 3.4.5, 1997-2002. <http://talana.linguist.jussieu.fr/~lionel/demo/demo-xlfg.html>
- DALYRMPLE M., KAPLAN R. M., MAXWELL III J. T., ZAENEN, A., ed: *Formal Issues in Lexical-Functional Grammar*. Stanford: Center for the Study of Language and Information CSLI, 1995.
- FALK, YEHUDA. *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. Lecture Notes No 126. Stanford: CSLI, 2001.
- FRANK, ANETTE. Syntax and Morphology of Tense in LFG. // *Proceedings of the LFG00 Conference*. <http://www.xrce.xerox.com/people/frank/papers.html>
- KAPLAN, RONALD M; BUTT, MIRIAM. The Morphology-Syntax Interface in LFG. (Abstract)// *Proceedings of the LFG02 Conference* (<http://csli-publications.stanford.edu/lfg02kaplanbutt-abs.html>)
- KING, TRACY HOLLOWAY. *Configuring Topic and Focus in Russian*. Stanford: Center for the Study of Language and Information CSLI, 1995.
- NEIDLE, CAROL. Lexical Functional Grammar. // *Encyclopedia of Language and Linguistics*. New York: Pergamon Press, 1994. P. 2147-2153. (<http://www.bu.edu/asllrp/neidle-lfg.pdf>)
- ROSEN, VICTORIA; ZAENEN, ANNIE. Grammar Writing in LFG: Introduction. // *Proceedings of the LFG99 Conference*. CSLI Online Publications. (<http://www2.parc.com/istl/groups/nlft/>)
- SADLER, LOUISA. New Developments in Lexical-Functional Grammar. // *Concise Encyclopedia of Syntactic Theories*. Brown, Keitsh; Miller, Jim, eds. Elsevier Science, Oxford. (<http://www.coli.uni-sb.de/~hansu/sadler.pdf>)
- SELLS, PETER. *Lectures on Contemporary Syntactic Theories*. Stanford: Center for the Study of Language and Information CSLI, Lecture Notes No 3, 1985.
- SHIEBER, STUART. M. *An Introduction to Unification-Based Approaches to Grammar*. Stanford: Center for the Study of Language and Information CSLI, 1986.
- SELJAN, SANJA. Unifikacijske gramatike kao okvir za leksicko-funkcionalnu gramatiku (LFG)// *Suvremena lingvistika* Sv. 1/2 Br. 47/48. Zagreb: Hrvatsko filološko društvo, 1999. P. 181-193.
- TALLERMAN, MAGGIE. *Understanding Syntax*. London: Arnold, 1998. <http://grid.let.rug.nl/~vannoord/papers/coling90/coling90.html>
- VINCENT, NIGEL; BÖRJARS KERSTI. Suppletion and syntactic theory. // *Proceedings of the LFG96 Conference*. CSLI Online Publications. (<http://lings.ln.man.ac.uk/Info/staff/KEB/Papers/Grenoble/Grenoble.html>)
- WESCOAT, MICHAEL T. *Practical Instructions for Working with the Formalisms of Lexical Functional Grammar*. Online University of Essex. (<http://www.fb10.uni-bremen.de/linguistik/khwagner/lfg/pdf/wescoat.pdf>)