Computer Science Journal of Moldova, vol.**14**, no.1(40), 2006

Local and Global Parsing with Functional (F)X-bar Theory and SCD Linguistic Strategy (I.)

Part I. FX-bar Schemes and Theory. Local and Global FX-bar Projections

Neculai Curteanu

Abstract

This paper surveys latest developments of SCD (Segmentation-Cohesion-Dependency) linguistic strategy, with its basic components: FX-bar theory with local and (two extensions to) global structures, the hierarchy graph of SCD marker classes, and improved versions of SCD algorithms for segmentation and parsing of local and global text structures. Briefly, **Part I** brings theoretical support (predicational feature and semantic diathesis) for handing down the predication from syntactic to lexical level, introduces the new local / global FX-bar schemes (graphs) for clause-level and discourse-level, the (global extension of) dependency graph for SCD marker classes, the problem of (direct and inverse) local FX-bar projection of the verbal group (verbal complex), and the FX-bar global projections, with the special case of sub-clausal discourse segments. **Part II** discusses the implications of the functional generativity concept for local and global markers, with a novel understanding on the taxonomy of text parsing algorithms, specifies the SCD marker classes, both at clause and discourse level, and presents (variants of) SCD local and global segmentation / parsing algorithms, along with their latest running results.

Notice. This is a paper in two parts, preserving a unitary numbering of the sections, and the unitary set and system of references along both parts.

1 Introduction: Basic Notions and Assumptions

This is a survey paper of the results, both theoretical and implementation aspects, concerning the latest form of *functional* X-bar (FX-bar) schemes (actually, graphs) and theory for *local* and *global* text structures, with a special accent on the problem of FX-bar (direct and inverse) projections of verbal group (verbal complex), and of the *SCD* (Segmentation-Cohesion-Dependency) linguistic theory and *segmentation / parsing strategy* for *local* (clause-level) and *global* (inter-clause and discourse) text structures. The paper is structured in two main parts, following the two main topics mentioned above.

In what follows, we shall try to specify, as much as possible, the basic notions we are working with. Within several papers and an evolution of the basic ideas along almost two decades [7], [9], [11], [16], [10], [17], the *SCD* (Segmentation-Cohesion-Dependency) linguistic strategy synthesized the following (more important) concepts and assumptions.

SCD considers four *major lexical categories* (and their functional projections within the FX-bar theory): *the Noun* (N) and *the Verb* (V) are the only lexical categories that have their own lexical (non-referential) meaning, and they are also saturated (representing their own semantic heads). Two other lexical categories play a central role in the syntactic organization of the *functional* X-bar (FX-bar) general schemes [11], [16].

The Adjective (Adj) has its own (auto-semantic) meaning but it is not a saturated lexical category, since it represents a modifier function to be applied to its intrinsic referentially nominal category, *i.e.* Adj is a modifier function that requires an N-type argument head. The pronominal adjective has a similar interpretation.

The Adverb (Adv) plays the role of V modifier, role similar to the one the Adj is playing for its N head. Often we denoted the modifier categories of Adj and Adv simply by A. It is important that this category is not confounded with the notation of A (Argument) positions (or A-bar, for non-argument positions), a common representation in classical linguistic theories. A special question is whether there exist properly *predicational adverbs*, as adjectives do ("*predicational*" feature in the sense of [15], often called *deverbal* property). It seems (at least for Romanian) that such adverbs do not exist properly. The first and most feasible explanation would be that the two predicational features of the verb and adverb would interfere, being too 'close' to each other. This is not a completely satisfying justification since both predicational noun and its adjective pair may coexist naturally!

These *four major* lexical categories are important because they may be endowed with *two* essential lexical-semantics features of the *local* (*i.e. clause* level, which is also the *predicational* level) syntactic-semantic structures of language organization: *Tense* and *Predication* features.

The feature *TENSe* (*Time*) may receive at lexical (syntactic) or phrase (analytical) level the values *FINIte* or *NonFINite* as well as various analytical combined values of tense and aspect for the temporal forms of the *verbal complex* [28], [30], [2], [23], [18]. The *FINIte* value of the feature TENS, for each of the four major (lexical and) syntactic categories, is borne at the lexical level or inherited from the lexical level by the verbal complex (to what traditionally is called *predicate*). For the structure of Verbal Complex in [28], [2] etc. we shall continue to use the term "Verb Group" (abbreviated VG), in order to remain consistent with the notions, theoretical and computational approaches of the functional FX-bar theory and SCD linguistic strategy. Both correspond, in a great measure, to the concept of *verbal predicate* in classical grammar.

V is the only *chosen* category for which the feature *TENSe* may receive its value *FINIte.* The other major categories, N and A (Adj, Adv), receive the value *NonFINite.* These values of the feature *TENSe* involve the construction of the local syntactic structures: the Noun Group (NG), which is the classical NP with a single nominal head, VG (the Verbal Complex, already referred), and the finite and non-finite clauses.

1.1 Classical and Lexical Predications

The feature that we called *Predicationality*, borne at the *lexical* (even *lexicon*) level by the major lexical categories N, V, A, corresponds to what in the literature is called (more frequently, among other labels) the *deverbal* property, or *deverbality*, of these categories. For an extended survey and analysis of the notion and its syntactic-semantic consequences, see [15]. We avoid the term *deverbality* because its meaning is *not necessarily specific* to V*s* since this essential lexical feature is *equally shared* by N*s* and A*s*. Moreover, there are (classes of) verbs which do not bear this property, *e.g.* the *copulative* ones. The feature of *Predicationality* is assigned to those finite or non-finite V*s*, N*s* (often called *nominalizations*), and A*s*, whose meaning involves a *process* event or a *process* name. We abbreviated this feature as PRED(dication)F(eature), with two main values, PROC(cess) and STAT(e) (or EXIST).

The classical notion of *predication* is known to be the pair (*Subject*, *Predicate*), an essentially *syntactic concept* meant to support the finite clause (proposition) structure. The predicate, either synthetic or analytic, encloses both *process* verbs and *state* verbs (the latter case for the nominal predicate) indiscernibly, despite the fact that only process (predicational) verbs entail an argument-based syntactic distribution, corresponding to a proper *valence.* Furthermore, the feature of *predicationality* (or *deverbality*) is equally shared not only by process verbs but also by nominals N*s* and modifiers A*s* that are (in term of lexical semantics) siblings of the corresponding predicational verbs, these non-verbal categories having a *similar syntactic distribution* of arguments, with the same valence as their predicational, verbal counterparts.

Thus, the feature of *predicationality* (being a lexical semantics quality) is not necessarily related to the predicate (which is a syntactic construction): in the nominal predicate, the copulative verb is not a predicational one. The same goes for the auxiliaries incorporated within the VG (verbal group, or verbal complex) whose tense is based on compound syntactic constructions. This does not exclude, in the nominal predicate, that the predicative nominal (as semantic head of the construction) bears the feature of predicationality. *E.g.*, the predicative nominals '*explanation*', '*marking*', '*receiving*' etc. (which are *predicational nouns*) in the nominal predicates of the clauses "*This is John's explanation* (*marking*, *receiving*, ...) *of the notion* ...".

These reasons support the idea of *handing down* the notion of *predication* from its classical, *syntactic* level to the *lexical*, word level of *representation* and *analysis*. The lexical semantics feature of *predicationality* (PREDF) has sometimes a contextual usefulness since the same word may, or may not, bear the feature PREDF, thus the process meaning depending on its contextual use. For instance, the noun "*building*" in languages like English, French, Romanian, may have both the meaning of a process, with [PREDF +] (or simply, PREDF), and the meaning of an object (in this case, the corresponding process result), with [PREDF –] (or STAT, or EXIST, or simply NPREDF values, see also [11], [12], [16]).

2 FX-bar Schemes for Local and Global Text Structures

2.1 Local (Clause-Level) Text Structures and FX-bar Projections

We pointed out within the SCD (Segmentation-Cohesion-Dependency) linguistic strategy that the natural language (NL) text is constructed from *local* and *global structures*. We consider *local structures* those structures that build a single finite-clause or a single (finite or non-finite) lexical predication (including both), in sum, finite or non-finite sub-clause and clause-level structures, while *global structures* represent inter-clausal or discourse level.

Thus a *local structure* is one of the following FX-bar structures: **(a)** single- (or multiple-) head *noun phrase*, together with its (their) FX-bar linguistic projection(s) (the single-headed noun phrase is called *noun group* NG in SCD); **(b)** single- (or multiple-) head *adjective phrase*, with its (their) FX-bar linguistic projection(s); **(c)** *finite verbal group* [7], [10], [11], known also under the label of *verbal complex* [28],

[29], [2], with its FX-bar projection elements (corresponding to what is also known to be the *verbal predicate*, either the *synthetic* or *analytic* one [20]; **(d)** *non-finite* VG, whose head is a non-finite V, bearing or not the predicational (deverbal) feature, and whose FX-bar projection is similar to that of the finite VG (verbal complex); **(e)** *finite clause*, viewed as the FX-bar projection of a finite VG; **(f)** *non-finite clause*, whose head is a *lexical* but *non-finite predicational* category (which can be a *predicational* but *non-finite* V, a *predicational* N, or a *predicational* Adj), together with its FX-bar projection.

Specif (or Spec) is also postulated in SCD to be a *functional* category bearing *quantificational features* at the lexical level (in particular, the negation at the X1 level), including (lexical or non-lexical) (in)definiteness, thus overlapping sometimes on the X1-marker functional features such as agreement.

The agreement (functional) relations are essential for what is called (*local*, *syntactic*) *cohesion* within SCD strategy: X0-Modif and X0-Specif agreement at the X1-level, Head-Subj and Compl-Pron$_{Emph}$ (Emphatic Pronoun) agreement at the X2-level etc. These kinds of (agreement, reference, and co-reference) *local cohesion* relations are responsible for a large category of *local dependencies*, including 'long-distance' dependencies. *Global cohesion* in the sense of [25], representing a chain of co-references for the same individual, is the discourse-level counterpart of a similar set of syntactic devices, but at the global level of text.

The FX-bar scheme for *local text structures* (including VG) is enclosed into the *line-bordered* (common) *part* of the figures 2.1. and 2.2. that present the *clause-level*, respectively *discourse-level global* FX-bar schemes.

2.2 Two Types of Global Text Structures

Global structures could be classified into (at least) two main categories: **(1)** There exist global structures built from *finite-clauses* or *lexical*

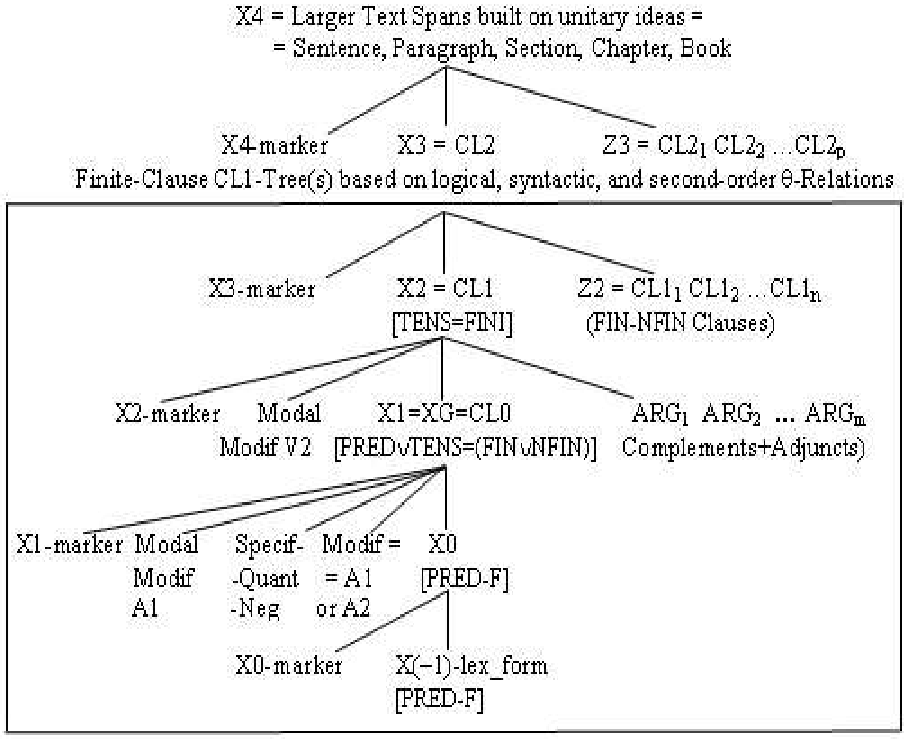

X4 = Larger Text Spans built on unitary ideas =
= Sentence, Paragraph, Section, Chapter, Book
X4-marker
X3 = CL2
Z3 = CL2$_1$ CL2$_2$ ...CL2$_p$
Finite-Clause CL1-Tree(s) based on logical, syntactic, and second-order θ-Relations
X3-marker
X2 = CL1
[TENS=FINI]
Z2 = CL1$_1$ CL1$_2$ ...CL1$_n$
(FIN-NFIN Clauses)
X2-marker
Modal
Modif V2
X1=XG=CL0
[PRED∨TENS=(FIN∨NFIN)]
ARG$_1$ ARG$_2$ ... ARG$_m$
Complements+Adjuncts)
X1-marker
Modal
Modif
A1
Specif-
-Quant
-Neg
Modif =
= A1
or A2
X0
[PRED-F]
X0-marker
X(−1)-lex_form
[PRED-F]

Figure 2.1. FX-bar scheme for clause-level local (and global) structures

predications (including both) using *logical* operators, *syntactic* operators (*e.g.* for the *relative clause*), and *second-order theta-relations* (*i.e. second-order predicational* relations, *e.g.* for the so-called subjective, predicative, direct-completive clauses etc.). **(2)** The usual *clause-based global text structures* are the sentence, paragraph, section, chapter etc.

The *clause-level global structures* correspond to the general FX-bar scheme whose elementary constructive element is the finite-clause (Fig. 2.1 above).

There exist *global structures* whose constructive bricks are not necessarily the finite-clause but the rhetorical *discourse-segment* of the RST discourse theory [24], [25], [26]. The FX-bar general scheme extension to RST *discourse-segment global structures* is presented in Fig. 2.2.

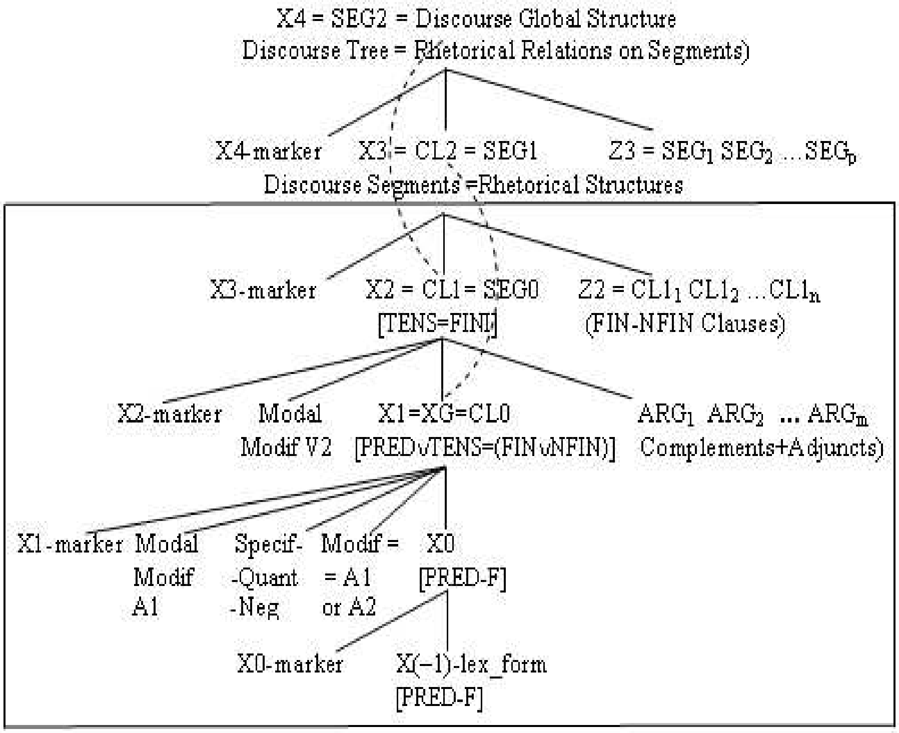


Figure 2.2. Discourse-level FX-bar (DFX-bar) Scheme

Remark. Dashed lines represent the special cases when a discourse segment is a proper subclause span and when a discourse segment splits a clause.

In general, a RST discourse segment is comprised of one or several finite-clauses. Actually, there exist an intricate relationship between the RST discourse segment and the finite clause, explored in [16]. Briefly, we have underlined that there exist sub-clausal discourse segments (*e.g.* a discourse segment constituted from a single NG), or that a discourse marker may split up a (finite) clause into text spans that belong to distinct discourse segments.

Significant elements involved by the *new linguistic projections* incorporate the discursive functionality within the currently proposed DFX-bar scheme (Fig. 2.2), while the categories and structures specified by the projection principles in the 'old', *local*, *clause-level* FX-bar scheme and theory remain the same (the bordered part in the figures 2.1 and 2.2).

2.3 FX-bar (Classes of) Markers and Their Graph-Graph Hierarchy

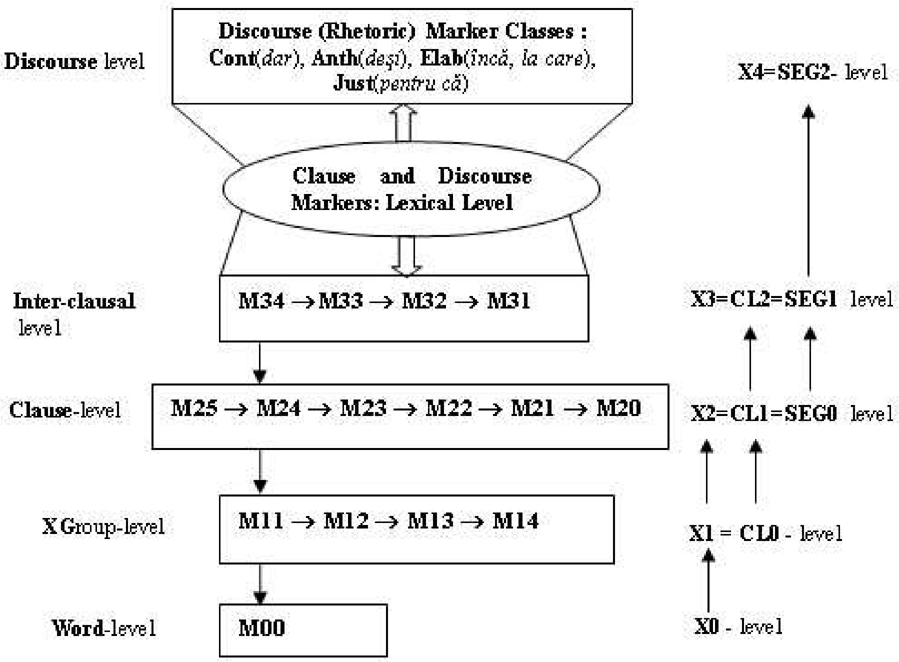


Figure 2.3. The hierarchy graph of SCD marker classes for local and global structures ([17])

One may naturally continue with relational (actually, functional) lexical (overt) categories, *e.g.* clause-level and discourse collocations (cue phrases, connectors, etc. called *clause* and *discourse markers*), but also lexically empty (covert) functional categories, such as T(ense) (or INFL) in [3], [4], [5], or the intrinsic presence of *predicational* (actually, *predicationality*) *feature* ascribed to a lexical category (and inherited by the XG phrase where that category is embedded in) [15] etc.

The device used to attain this goal is what we called *classes of functional markers*, together with their corresponding *hierarchies* [8], [9], [35], [17]. The dependency graph of the hierarchy of SCD marker classes (Xn-marker in Fig. 2.3) is represented in [14], [17]. The *explicit description* of the marker classes that are used within the FX-bar schemes in Fig. 2.1 and 2.2, and in the dependency graph in Fig. 2.3 is exposed in Section 7, Part II of the paper.

3 Local FX-bar Projections

3.1 Verbal Group Kernel and the Predicational Feature

The verbal group (VG), as XG structure in the BAR = 1 projection level of the FX-bar scheme, contains a *semantic head* verb, *around* which one can find pronouns (only in unaccentuated forms, *i.e.* clitics), special adverbs, auxiliaries, modal verbs (or adverbs), negation. VG is also better known under the label of *verbal complex* (see [28], [29], [2], [23]), and constitutes what is traditionally called *verbal predicate* for the classical clause (proposition). The VG Kernel (VGK) was initially introduced in [17, p.175] (under the name of *default verbal kernel*), and represents a basic substructure in the VG parsing. The *typical difference* between VG and VGK is that VGK is missing the *proper adverb* of VG (that may syntactically commute with VGK to accomplish the VG).

Examples of VGK*s* [17] (VGK is represented in parantheses, included in VG; unaccentuated pronouns (clitics) in VGK are in *italics*):

" nu că (nu *mi-l* va mai şi plăti) greu; (nu-*i* cunoşteam); (*li se* cereau) ; (*îşi* mai recăpătase) ; (Ai consultat) ; (ar fi simţit) ; (*i se* aşternea) ; (să *se* întâmple) ; (nu *se* putea abţine); (n-*o* putea lua); (Nu *i*-ar fi trecut); (să poată afla); (să *te* intimideze); (să *vă* văd lucrând) ".

VG may be seen as the shell of VGK, while the contents of VGK may be interpreted as the *clause-shadow* (of the regular clause) that projects itself onto the clause, as well as representing the projection(s) of the lexical-semantic head bearing the *predicationality feature* (*e.g.* [15]), using *diathesis transformations* and *semantic diathesis functions* associated with semantic restrictions on predication arguments (see [29], [28], [2], [23]).

A rightful observation in [2] is that VG provides both an *outside* (**nu**$_1$) *negation* and an *inside* (**nu**$_2$) *negation* (*e.g.* "***nu***$_1$ *să* ***nu***$_2$ *te duci*"), which can be interpreted as outside (VG) and inside (VGK) *quantifiers*. Similarly, there exist as VG *modifiers* the (VGK) *special, inside adverbs* ("cam", "mai", "prea", "şi", "tot"), and the proper, VG *outside adverbs* ("*nu*$_1$ *să nu*$_2$ *te* ***tot***$_2$ *duci* ***imediat***$_1$"). The structure of VGK as the "inside" of VG, with a *syntactic head* (the auxiliary bearing the number and person, when lexically present) and a *semantic one* (the predicational verb), with clitics 'inside' (and arguments 'outside') VGK playing an essential role in the development of the *lexical predication* should be further explored, both in linguistic theory and parsing.

In a *verbal group* (VG), the "positive" feature values such as PROC and FINI are *inherited* from the tensed V head by the whole VG phrase, or may be *cumulatively acquired* through morpho-syntactic FX-bar projection.

Somehow similarly (preserving proportions) to the A. Joshi's well-known *tree adjoining grammar* (TAG) and *lexicalized* TAG (LTAG) [27], SCD strategy may also be seen as a *theory* of (D)FX-bar *scheme* (thus *tree*) *checking and adjoining.* Since in LTAG one considers the *initial trees* to be of the form 'functor-arguments', thus one begins in phrase generation with a clause shell, our VGK, whose structure is a *clause-shadow*, may constitute a substantiation argument for *initial trees* in lexicalized TAGs.

In [15] we discussed a suitable taxonomy for classical predications, involving the classical predicates (VG or verbal complex), based on the lexical property of predicationality PREDF, in agreement also with the extensional / intensional logical representations of these structures.

A typical example of the SCD predicational taxonomy is given by the two main categories of common nouns: **(i)** non-*predicational nouns*, corresponding to *existential-type*, *object-denoting*, *non-event individuals*, whose predicational feature PREDF value is EXIST (*e.g.* [Eng: *student, table*; Rom: *elev-student, masă*]), and whose functional representation in *extensional logic* is done by predicates depending on a single, *extensional variable*: *student*(X), *table*(X) etc.

Our interest is however in **(ii)** *predicational nouns* (often called *nominalizations*), whose predicational PREDF feature value is PROC, *e.g.* [Eng: *meeting, envy, marking,* etc.; Rom: *întâlnire*, *invidie*, *marcare*, etc.], whose functional representations depend on several *intensional variables*, *e.g.* *întâlnire*(x, y, ...), *invidie*(x, y, ...), *marcare*(x, y), *donaţie*(x, y, z) etc. Proper nouns and/or personifications are encoded either as constants or variables of extensional nature on which the above extensional / intensional predicates are applied. Other examples: the common nouns *car* and *man* are non-predicational individuals, represented extensionally as *car*(X) and *man*(X), the adjective *red* is a basic, also *extensional* predicate *red*(X), while *leaving* is a *predicational* (process-event) *nominal* (also called *nominalization*) which is represented as an intensional (unsaturated) predicate *leaving*(x, y), with x and y as *intensional* arguments.

[Eng: *boy, pencil*; Rom: *băiat, pix*] PREDF = EXIST, and TENS=NFIN;

[Eng: *attempt, showing, proved*; Rom: *încercare, arătând, demonstrat*]
PREDF = PROC, and TENS := NFIN;

[Eng: *are*; Rom: *sunt*] PREDF = EXIST, and TENS := FINI;

[Eng: *gives*; Rom: *dă*] PREDF = PROC, and TENS := FINI.

The *predicational nouns* are typical non-verbal categories whose distributional behaviour is perfectly similar to their verbal counterparts included in VG*s*.

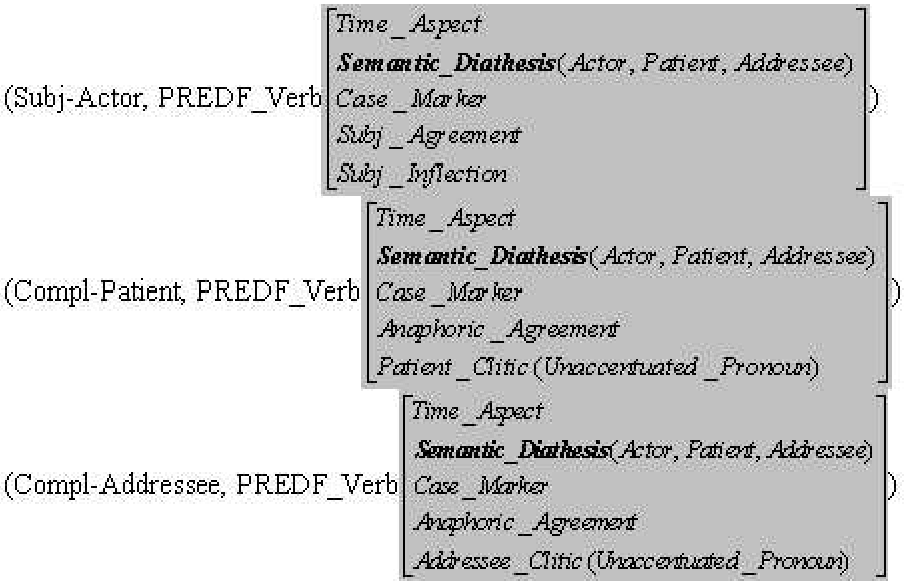
3.2 From Syntactic to Lexical Predication

Without coming into details (see [18]), the classical predication pair (Subject, Predicate) can be viewed as just *one of the facets* of the VG (verbal complex) whose semantic head bears PREDF, the other ones, equally righted as "classical predications", being instantiated by the predicational verb (lemmatized form), endowed with clitic(s) as affixed inflexion(s), which are obligatory present when their valence-based arguments are of personalized semantic nature and optionally present otherwise, doubled or not by the corresponding valence-commanded arguments. Thus, the classical predication pair corresponds to the subject theta-role of "actor" or "actant", while the other "classical" predications associate, valence-driven, the theta-roles of "patient" and/or "receiver" and/or "addressee" to semantic arguments (but not adjuncts!). All these are commanded (or not) by the presence (or absence), at the *lexical level*, of the PREDF feature assigned to the semantic head in VG (verbal complex).

Thus, in a *first move*, the classical predication pair (Subject, Predicate) should be reduced to the pair (Subject, PREDF_verb) corresponding to the *theta*-role of "actor" or "actant" in the valence-driven SUBCAT vector (with 1 to 3) semantic arguments. It is important to specify that there exist normally at least two SUBCAT lists: $\text{SUBCAT}_{oblic\_order}$, containing the syntactic arguments of the PREDF_verb, in the order of increasing obliqueness, and $\text{SUBCAT}_{theta\_order}$, enclosing the arguments in the *theta*-order (or *systemic* order) for the valence-based arguments of PREDF_verb. Usually, (only) *for the active voice* and a normal semantics of predicationality, these arguments should coincide.

In a *second move* the following similar "classical" predications (see Figure 3.2) are added to this predication, equally righted in the *theta*-semantics.

These are the new 'traditional' predications, with their real engine, *viz.* the predicational feature PREDF, installed on the verb head of the verbal group VG (verbal complex). Similarly, non-finite forms of PREDF verbs may be associated to those N*s* (called nominalizations) and/or As that bear the feature PREDF.



$$\text{(Subj-Actor, PREDF\_Verb}\begin{bmatrix} \mathit{Time\_Aspect} \\ \boldsymbol{\mathit{Semantic\_Diathesis}}(\mathit{Actor}, \mathit{Patient}, \mathit{Addressee}) \\ \mathit{Case\_Marker} \\ \mathit{Subj\_Agreement} \\ \mathit{Subj\_Inflection} \end{bmatrix}\text{)}$$

$$\text{(Compl-Patient, PREDF\_Verb}\begin{bmatrix} \mathit{Time\_Aspect} \\ \boldsymbol{\mathit{Semantic\_Diathesis}}(\mathit{Actor}, \mathit{Patient}, \mathit{Addressee}) \\ \mathit{Case\_Marker} \\ \mathit{Anaphoric\_Agreement} \\ \mathit{Patient\_Clitic}(\mathit{Unaccentuated\_Pronoun}) \end{bmatrix}\text{)}$$

$$\text{(Compl-Addressee, PREDF\_Verb}\begin{bmatrix} \mathit{Time\_Aspect} \\ \boldsymbol{\mathit{Semantic\_Diathesis}}(\mathit{Actor}, \mathit{Patient}, \mathit{Addressee}) \\ \mathit{Case\_Marker} \\ \mathit{Anaphoric\_Agreement} \\ \mathit{Addressee\_Clitic}(\mathit{Unaccentuated\_Pronoun}) \end{bmatrix}\text{)}$$

Figure 3.2. All the extended, valence-based 'classical' predications

In the 'classical' predications above, clitics may lack when the semantic arguments are of *non-person* or *non-animate* nature but are lexically present. This does not change the 'equivalence' of these newly devised valence-based predications. Such an interpretation of the VG structure has consequences in establishing the FX-bar (direct and inverse) VG projections (see the outlined solutions considered in the subsection 3.1 devoted to the problem of VG local structure and its FX-bar projections).

The problem of 'classical' *predication*(s) in HPSG [1], or the problem of the *special* role of the *subject* in the SUBCAT list of HPSG theory [34; Chap.9] are solved in the linguistic feature structures in Fig.3.2. above as follows: the feature *Semantic_Diathesis*(*Actor*, *Patient*, *Addressee*) is not an elementary (atomic) feature value but a *function*, whose input value is the VG shallow, *syntactic diathesis*, represented by the above mentioned $\mathrm{SUBCAT}_{oblic\_order}$, while the output value of the function is the VG *semantic diathesis*, *viz.* $\mathrm{SUBCAT}_{theta\_order}$ list. This solution forces the subject-*actor* and the subject-*least_oblique_element* (or grammatical subject) to take each one its own right place, in the right (possibly distinct) ordering.

In the parsing / generation processes, the *input value* of the function *Semantic_Diathesis* is represented by the tense and syntactic diathesis resulted from the VG shallow parsing. The *output value* of *Semantic_Diathesis* function is obtained from the lexicon, where the head verb (*predication*) meaning is represented by specific standard lists of semantic arguments corresponding to the valence of that specific predicational category, and the syntactic diathesis is transformed into a certain particular list of semantic arguments corresponding to the tense, diathesis, and predicational meaning of that (verb) category. [18] describes in detail the mechanism of *diathesis transformation* and *semantic diathesis* functions, defined to make operational the effective resolution of VGK direct and inverse FX-bar projections (see §3.3).

In Fig. 2.1.-2.2. of the FX-bar scheme for *local structures*, the local (single-event) levels X0-X1-X2 express the clause predication depending on basic, lexical categories, while the levels CL0-CL1-CL2 express logical or (second-order) predicational relations on simple clauses. The two global FX-bar schemes work in a (top-down and bottom-up) recursive manner, both in the analysis and generation tasks of the parser, in close relationship with SCD linguistic strategy, its marker classes and hierarchy, and its meta-algorithms of analysis-generation being exposed in [9], [10].

3.3 Direct and Inverse FX-bar Projections of VGK

In [18] we introduced *diathesis transformations* (*dt*) and *semantic diathesis* (*sd*) *functions* as useful tools in describing the lexical predication metamorphosis from syntactic (shallow) diathesis to semantic diathesis as a top-down and bottom-up movement, from text to lexicon and backwards. This process may also be understood as direct and inverse FX-bar projection procedures of VG (VGK) towards its (predicational) semantic head and to the clause, derived from the diathesis analysis (described as in [26]), stated as solutions to the following VG (VGK) FX-bar projection problems:

FX-bar(VG): The problem of direct FX-bar projection of VG: To show how the *clause-shadow* information (see above) incorporated into VG is (directly) FX-bar projected into a (finite or non-finite) regular clause.

FX-bar^{-1}(VGK): The problem of inverse FX-bar projection of VGK: To obtain an improved linguistic mechanism by which a predicational category (from the lexicon) is FX-bar projected on VG (VGK). This means to establish the FX-bar *inverse projection* FXbar^{-1}(VGK) for VGK (or VG), *i.e.* the morphologic-phonologic-syntactic-semantic restrictions on the (predicational) semantic head of VGK that are necessary (and sufficient) to retrieve the VG (VGK) *local* structure through (direct) FX-bar projection of its semantic head.

The inverse FX-bar projection associates to VGK a number of (virtual) semantic heads, corresponding to the meaning(s) of the lexical head entry, each semantic head observing the set of *diathesis transformations* (*dt*) and *semantic diathesis* (*sd*) *functions* and *values*, along with phonologic, lexical, morphologic, syntactic and semantic restrictions at lexical level on arguments, clitics, doubling etc.

This is the starting point in the process of *generation task*, when the first requirement is to generate one or several adequate VGs, satisfying the text *planning* restrictions. For clause analysis / generation, the parsed VG (as *clause-shadow*) or the obtained VG(s) is FX-bar projected into one (or more) finite or non-finite clause(s), with its (their) arguments, constructed lexically from diathesis computations and linguistic restrictions.

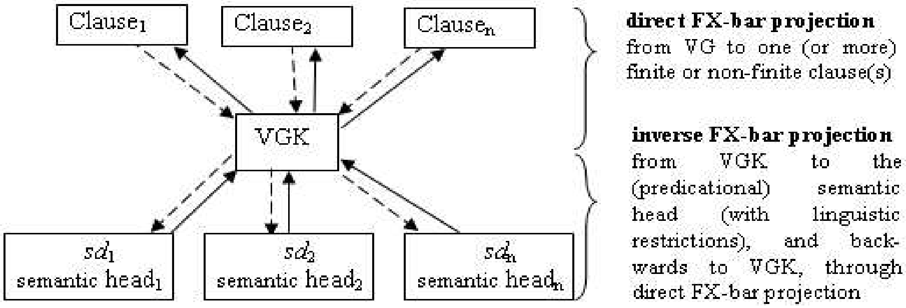

Clause$_1$
Clause$_2$
Clause$_n$
VGK
sd_1 semantic head$_1$
sd_2 semantic head$_2$
sd_n semantic head$_n$
direct FX-bar projection from VG to one (or more) finite or non-finite clause(s)
inverse FX-bar projection from VGK to the (predicational) semantic head (with linguistic restrictions), and backwards to VGK, through direct FX-bar projection

Figure 3.3. FX-bar projections of VGK, from text to lexicon and backwards

4 Global FX-bar Projections

4.1 Direct and Inverse Global FX-bar Projection

It is a common fact in the classical grammar to expand the thematic arguments (*theta*-roles) from inside the clause to the inter-clausal relations of the same type, using such labels as *subjective*, *predicative*, *direct-completive* etc. *clauses*. Such a clause tree, whose *inter-clausal* relations are based on *logical-type* operators (conjunction-disjunction, implication, conditional, concession, consecution-purpose, correlated operators, such as *if-then-else* etc.), (purely) syntactic-type relations (such as the *relative clause*), and *second-order theta-semantics* relations (as the above mentioned *theta*-role clauses) could be considered the *global linguistic projection* of the *saturated matrix (root) clause* of the *clause-level tree*. Other weaker (syntax or grammatical-oriented) semantics, together with node operations on the clause-tree may be taken into account.

A *similar problem* can be stated for the discourse segments, in particular for the *discourse tree* evolved from the RST inter-segment *rhetorical relations* [24]. Questions of theoretical and practical (computational) importance: which is the discourse projection nucleus for a resulted RST discourse tree, and what is the relationship between the corresponding clause and discourse segment trees?

In terms of discourse tree, we may state the following conjecture: a *text could be seen as* the “global” *projection of its discourse tree* (or of certain subtrees of the discourse tree). It is a kind of “summarization” of the text through its main rhetorical components (discourse segments), hierarchically organized as its discourse tree. A similar conjecture may be stated in terms of the corresponding *finite clause-level tree* (or certain subtrees), as a hierarchically organized tree (or graph) of the text enclosed events.

As one can see from Figures 4.1.1.–4. and examples Ex. 4.1.1.–2., the clause-level trees are not necessarily embedded into the corresponding discourse segment trees: in Ex. 4.1.1., a subclause phrase makes a unitary segment with clauses in the following sentence(s), and in Ex. 4.1.2., a subclausal phrase (a non-finite clause) is broken (detached) and adjoined to the next clause, making a discourse segment.

Making comprehensible the (global) projection function is important also from another (somehow surprising) point of view: *anaphora resolution. Referring* an individual (object or person), a process (event or existence), or a whole bunch of actions that corresponds to a larger text span is equivalent to equating the referee category as the value of the ‘*inverse*’ *of linguistic projection function* applied to the corresponding text. In other words, a phrase that points at a specific text span would be naturally associated with the *head* (or *nucleus*, kernel, projection tree, or another linguistic object) that is (locally or globally) projected into that text, thus with the value of the *projection function inverse* applied to that text. This perspective shows complementary facets of the linguistic projection mechanism and its specification.

Two simple examples may give a better idea of the approach we propose: “*Plecarea vânătorilor în munţii Călimani pe o vreme atât de rea a fost pe nepregătite.* Aceasta *le-a fost fatal.*” The demonstrative pronoun “*Aceasta*” refers to the whole previous sentence, and one could associate it with the sentence (and finite clause) predicate head “*a fost pe nepregătite*”, or even corroborate it with the predicational head “*plecarea*” of the enclosed non-finite clause. These phrases represent “*inverses*” of the projection function, applied to the whole sentence at the local, clause-level.

Another possible example is to associate the phrase that refers a whole story within a (larger) text span to the discourse or clause-dependency tree of that text, *i.e.* to the value of '*inverse*' *projection function*, applied to that text, at the global level. This correspondence relates the story reference expression to specially computed nodes and/or subtrees in the mentioned trees.

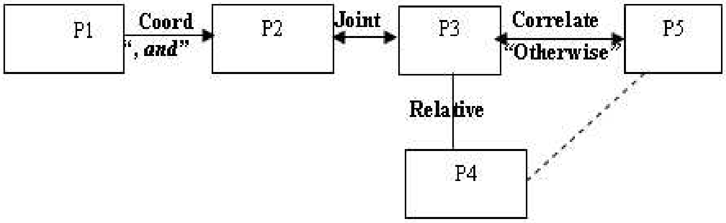


Figure 4.1.1. Inter-clause tree inherent to the segment tree of Ex. 4.1.1.

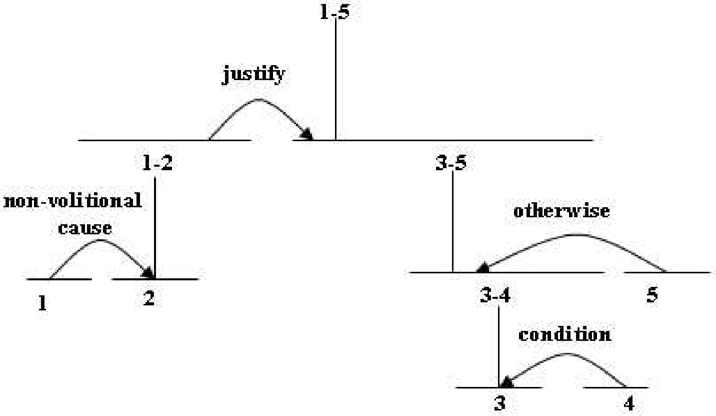


Figure 4.1.2. The RST segment tree for Ex. 4.1.1. [24; Fig. I-13]

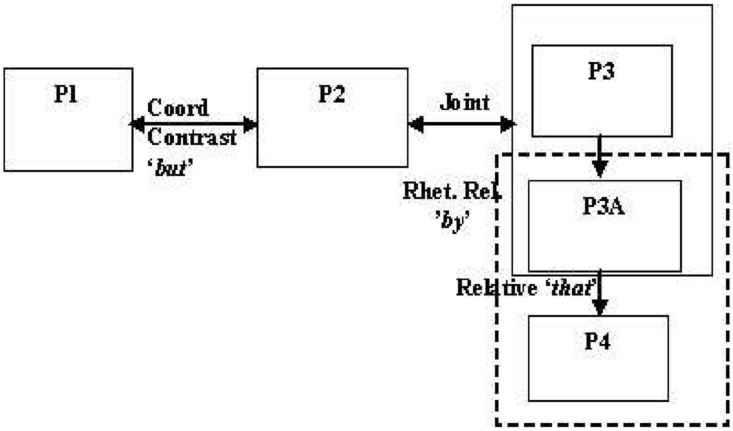


Figure 4.1.3. Clause-level tree inherent to the segment tree of Ex. 4.1.2.

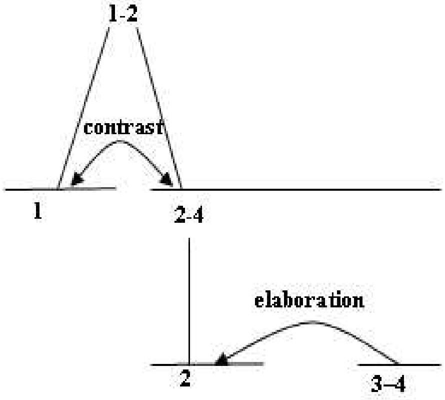


Figure 4.1.4. The RST segment tree for Ex. 4.1.2. [24; Fig. I-19]

The problems of *linguistic projection* at the *global level*, floors 3 and 4 in DFX-bar scheme (Fig. 2.2), are especially complex. Two (counter)examples show that an RST discourse segment is not necessarily the projection of its enclosed *saturated matrix clause,* as one would expect. Corroborated with the fact of subclausal discourse segments [14], [17], this gives the flavour for the difficulty of the problems for specifying the *discourse* (*global*) *projection function*, as well as *its* '*inverse*' one, *i.e. the nucleus* (or *head*) *structure* whose projected value is a certain (larger or smaller) text span.

Ex. 4.1.1. [24; p.68] **(1)** *Un nou număr din broşură este în curs de apariţie,* **(2)** *şi acest lucru înseamnă o şansă pentru noi propuneri de proiecte.* **(3A)** *Oricine* **(4)** *doreşte să actualizeze intrările în broşură* **(3B)** *ar trebui să aibă copia până la 1 Decembrie.* **(5)** *În caz contrar va fi utilizată intrarea existentă.*

[24] notices that, for the rhetorical relations **condition** and **otherwise**, their classical constructions for RST diagrams do not cover the text (similar to the classical programming) conditional "*If* A, *then* B. *Otherwise* C". These syntactic constructions receive a special attention in the latest version of SCD, falling under the (important) category of *correlated constructions* (and *clauses in correlation*) [17] (see Subsection 8.1, Part II).

Ex. 4.1.2. [24; p.76] **(1)** *Animalele se vindecă,* **(2)** *dar arborii se compartimentează.* **(3)** *Ei rezistă întreaga viaţă la răni şi infecţii* **(4)** *prin instalarea unor graniţe care rezistă la extinderea microorganismelor invadatoare.*

The figures 4.1.1-4.1.4 showing the inter-clausal relations and discourse trees for the texts in Ex.4.1.1–2. support our statements concerning the global projections at these clause and discourse levels.

4.2 A Special Case: Sub-Clausal Discourse Segments

The essential difference between Marcu's discourse segmentation [25], [26], and the SCD syntax-driven segmentation is the type of target structures that the two algorithms are looking for: Marcu's algorithm's objective is to obtain structures derived from RST rhetorical relations [25], while SCD's main purpose is to reveal the sentential, syntactic-semantic structures, at the sentence level, from syntactic category-headed phrases and non-finite clauses, to finite clauses and inter-clausal (syntactic and logical-semantic) relations. Between rhetorical relations and inter-clausal relations of syntactic-semantic nature there is a subtle, distinctive, however close relationship. The following examples from [25; Appendix A] illustrate some aspects of this situation:

Ex. 4.2.1. [25; Text A.4, p. 268] [*Every rule has exceptions,*] [*but the tragic and too common tableaux of hundreds or even thousands of people snake-lining up for any task with a paycheck illustrates a lack of jobs,*] [*not laziness.*]

Comments. The discourse segment [*not laziness.*] is actually a clause, defective of its (finite) predicate "*illustrates*".

Ex. 4.2.2. [25; Text A.6, p. 269] [*Cleaning agents on the burnished surface of the Ectype coating actually remove build-up from the head,*] [*while lubricating it at the same time.*]

Observation. The situation when an elementary discourse unit (EDU, or discourse segment) is properly embedded into a (finite) clause is very close to that when a discourse marker splits a finite clause into two spans, each span belonging to distinct EDU*s*. This does not necessarily mean that the two EDU*s* are both enclosed into the same finite clause; the most frequent situation is when a discourse segment tears a phrase from a clause and continues its span on the next clauses(s).

These examples bring further arguments that the relationship between the *discourse segment tree* and its underlying *clause-level tree* is an intricate interplay, falling both in the field of lexical semantics for *local structures* but especially in the area of clause-level and discursive semantics for *global structures* of the text organization. Segment tree projection function is closely related to the composition between the discourse marker semantics and the clause-level predications involved by the subsumed clauses. Until one would know more about the projection functions at global level, about the relationship between clause-dependency and discourse trees of a text span, and how the global projection functions and their inverses work, these issues remain still open and challenging problems.

5 Transitory Conclusions

The *phrase markers* play a fundamental role in delimiting the syntactic (and also, semantic) structures and establishing their dependencies. I emphasized this since the beginnings of SCD ([7] and earlier). One can see now a whole movement toward rediscovering the essential role of markers, especially on the *discourse* and higher levels of the text. The SCD linguistic strategy, in particular the local and global (D)FX-bar theory, is trying to use and to put to work not only the '*connectives*' of several types, '*cue phrases*', '*discourse markers*', etc. but the *whole palette* of markers, for *all* the (local and global) levels of analysis-generation of NL, from lexical to discourse ones, especially *in the syntax.* SCD makes a special effort to maximize the use of the lexical-semantics and syntactic means in discovering the logical-semantics and discursive structures of NL.

The main novel aspects that make the difference between (D)FX-bar schemes and previous X-bar type theories may be summarized as follows: **(a)** the two *global* FX-bar schemes that represent extensions to the *clause* and *discourse levels*, both enclosing an improved shape of the same *local* FX-bar *scheme*; **(b)** the *graph-based hierarchy* of the SCD *marker classes* that are used in the FX-bar general schemes, for the local and global levels; **(c)** theoretical arguments supporting a *lexical-level predication*, based on the *predicational feature* assignment to the major lexical categories N, V, A, since the lexicon description; **(d)** local and global FX-bar projection problems, with emphasis on the fundamental VG (and VGK) local structures; **(e)** The maximal use of the functional (predicational) and relational features of the local and global markers represent an adequate framework for defining the concept of *functional generative capacity* (in Part II), with interesting consequences on the design and taxonomy of local / global text segmentation / parsing algorithms.

DFX-bar scheme may be associated also to a *language-dependent automaton* (working similarly, however, for a large class of NL*s*) that starts with a sentence, receives *on-line* each word of it, and stops at the final punctuation sign. For adequate values of the parameters like *word* (*argument*) *ordering* and *projection direction* of the major categories and markers, the FX-bar scheme can properly represent the correct dependency of linguistic structures.

As a basic component of the SCD linguistic strategy [9], the *local and global* FX-bar *theory* may also be seen as a *procedural mechanism* providing a consistent set of principles and rules that ensure a sound functioning of the FX-bar *schemes*, from the lexicon to the discourse level organization of the NL analysis / generation processes. Continuing this perspective and paraphrasing A. Joshi's well-known *tree adjoining grammar* (TAG) [21], SCD strategy may further be understood as a *theory* of (D)FX-bar *schemes* (thus *tree*) *checking and adjoining* (see subsection 3.1). The same role of procedural mechanism for FX-bar scheme(s) is envisaged for the related but more general model of *Marcus contextual grammars* [33], as a down-to-language strategy putting to work (highly)-*contextual mechanisms* (such as SCD marker classes and dependency principles) for the NL phrase structure recognition and generation. These are just samples of the role that FX-bar theory can still really play within the NL theory and technology.

The global (D)FX-bar scheme exposed in Fig. 2.2 represents an essential extension to the global approach in the context of SCD linguistic strategy. Each of major lexical categories X = N, V, A, along with the grammatical category CL and the discourse category SEG, are projected (recursively) on *three* bar levels (BAR = 0, 1, 2), within *five* local-global levels of FX-bar linguistic projection process. All these structures, except the lexicon normalized X0-lex form, are functionally and/or relationally "marked", through multiple applications, by the *four*-level local / global markers (on the first level of the hierarchy, followed by other sub-hierarchies) whose classes are better specified within the SCD linguistic strategy (Section 7, Part II).

Functional properties of the (predicational or relational) categories can be assigned even from the lexicon level, but the semantic and/or pragmatic context may entail temporarily loosing or gaining such a quality. This is also true for phrases and collocations resulting from the lexical analysis. Discovering and pointing out the *functional* (*functorial*) and *relational properties* of the words and phrases is an essential task of the NL *parsing* (analysis and generation) processes; this is specific not only to SCD linguistic strategy, but also to *principle-based parsing* strategies, *e.g. rhetorical parsing* [25]. The proposed FX-bar schemes, consolidating the basic ideas of AX-bar schemes [7], [9] and FX-bar theory [11], [12], [16], provide both a theoretical and practical tool for local and global parsing / generation text processing tasks.

A central issue for obtaining a solution to the *direct* and *inverse* FX-bar *projection problems* of VG (VGK) consists in defining and computing *diathesis transformations* and *semantic diathesis* functions, showing that these function values may characterize the way VG structure is (reversely) FX-bar projected into its (predicational) semantic head, as well as (directly) FX-bar projected, on the lines of its semantic head meaning(s), into the corresponding clause(s). This mechanism is described in [18], supporting a better understanding of *lexical predication* anatomy and functioning.

References

[1] Barbu, Ana-Maria; Emil Ionescu (1996): *Contemporary grammatical theories*: *grammars of the phrasal head.* in Limba româna, no. 45, (1-6) pp. 31–55 (in Romanian).

[2] Barbu, Ana-Maria (1999): *The Verbal Complex.* Studii si Cercetari Lingvistice, L, no.1, Bucureşti, p. 39–84 (In Romanian).

[3] Chomsky, Noam (1981): *Lectures on Government and Binding.* Foris, Dordrecht.

[4] Chomsky, Noam (1986): *Barriers.* The MIT Press, Cambridge.

[5] Chomsky, Noam (1995): *The Minimalist Program.* The MIT Press, Cambridge, Massachusetts.

[6] Cristea, D., O. Postolache, I. Pistol (2005): *Summarisation through Discourse Structure.* In Proceedings of CiCling 2005, Springer LNSC, vol. 3406.

[7] Curteanu, Neculai (1988): *Augmented X-bar Schemes.* COLING'88 Proceedings, Budapest, pp. 130–132.

[8] Curteanu, Neculai (1990): *A Marker-Hierarchy-based Approach Supporting the SCD Parsing Strategy.* Research Report no. 18, Institute of Technical Cybernetics, Bratislava, Slovak Republik.

[9] Curteanu, Neculai (1994): *From Morphology to Discourse Through Marker Structures in the SCD Parsing Strategy. A Marker-Hierarchy Based Approach.* Language and Cybernetics, Akademia Libroservo, Prague, pp. 61–73.

[10] Curteanu, Neculai; G. Holban (1996): *SCD Linguistic Strategy Applied to the Analysis and Generation of Romanian.* In (Dan Tufiş, Ed.) Language and Technology, Romanian Academy, Bucharest, pp. 169–176 (in Romanian).

[11] Curteanu, Neculai (2002): *Elements of a Functional X-bar Theory Within the SCD Linguistic Strategy*, ECIT2002 Conference, Iaşi, România.

[12] Curteanu, Neculai (2003): *Towards a Functional X-bar Theory.* In the volume "The Romanian Language in the Informational Society", Dan Tufiş, F. Filip (Eds.), Edited by the Romanian Academy, Research Institute for Artificial Intelligence, Bucharest, pp. 51–86 (in Romanian).

[13] Curteanu, Neculai; D. Gâlea; C. Linteş (2003): *Segmentation Algorithms for Clause-Type Textual Units.* In the volume "The Romanian Language in the Informational Society", Dan Tufiş, F. Filip (Eds.), Edited by the Romanian Academy, Research Institute for Artificial Intelligence, Bucharest, pp. 165–190 (in Romanian).

[14] Curteanu, Neculai; D. Gâlea; C. Butnariu; C. Bolea (2004): *Marcu's Clause-like Discourse Segmentation Algorithm and SCD Clause Segmentation-based Parsing*, In the volume "Intelligent Systems" (Ed. Horia-Nicolai Teodorescu). Selected Papers from ECIT2004 Conference, Iaşi, România, pp. 59–86.

[15] Curteanu, Neculai (2003-2004): *Contrastive Meanings of the Terms* "Predicative" *and* "Predicational" *in Various Linguistic Theories* (I, II). Computer Science Journal of Moldova (R. Moldova), Vol. 11, No. 3(33), 2003 (I); Vol. 12, No. 1(34), 2004 (II).

[16] Curteanu, Neculai (2005): *Functional FX-bar Theory Extended to Discourse (Rhetorical) Structures.* In 'Intelligent Systems' Conference Volume, H.-N. Teodorescu *et al.* (Editors), Performantica Press, Iaşi (Romania), pp. 169–182.

[17] Curteanu, Neculai; E. Zlavog; C. Bolea (2005): *Sentence-Based and Discourse Segmentation / Parsing with SCD Linguistic Strategy.* In 'Intelligent Systems' Conference Volume, H.-N. Teodorescu *et al.* (Editors), Performantica Press, Iaşi (Romania), pp. 153–168.

[18] Curteanu, Neculai; Diana Trandabăţ (2006): *Functional (F)X-bar Projections for Local and Global Text Structures. The Anatomy of Predication.* Revue Roumaine de Linguistique, Bucharest (to appear).

[19] Dobrovie-Sorin, Carmen (1994): *The syntax of Romanian. Comparative Studies.* Berlin: Mouton de Gruyter.

[20] Irimia, Dumitru (1997): *The Morphosyntax of the Romanian Verb.* The Editorial House of the "Al. I. Cuza" Iaşi University (in Romanian).

[21] Joshi, Aravind K. and Ives Schabes (1997): *Tree Adjoining Grammars.* In "Handbook of Formal Languages and Automata" (A. Salomaa *et al.*, Eds.), Vol. 3, Heidelberg, Springer-Verlag.

[22] Kornai, András, Geoffrey Pullum (1990): *The X-bar Theory of Phrase Structure*, Language, Vol. 66, No. 1, pp. 24–50.

[23] Legendre, Géraldine (1999): *Optimal Romanian clitics: a cross-linguistic perspective.* In: V. Motapanyane (Ed.) Comparative Studies in Romanian Syntax. HAG, The Hague.

[24] Mann, William, Sandra Thompson (1988): *Rhetorical Structure Theory: A Theory of Text Organization.* Research Report RS-87-190, Information Sciences Institute, University of Southern California, Marina del Rey, California, 80 pp.

[25] Marcu, Daniel (1997): *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*, Ph.D. Thesis, Univ. of Toronto, Canada, pp. 341.

[26] Marcu, Daniel (2000): *The Theory and Practice of Discourse Parsing and Summarization.* The MIT Press.

[27] Miller, Philip (1999): *Strong Generative Capacity. The Semantics of Linguistic Formalism.* CSLI Publications, Stanford, California, 1999.

[28] Monachesi, Paola (1998): *The Morphosyntax of Romanian Cliticization.* In: P.-A. Coppen *et al.* (Eds.), Proceedings of Computational Linguistics in The Netherlands 1997, pp. 99–118, Amsterdam-Atlanta: Rodopi.

[29] Monachesi, Paola (2000): *Clitic placement in the Romanian verbal complex.* In: B. Gerlach and J. Grijzenhout (eds.) Clitics in phonology, morphology, and syntax. Linguistik Aktuell. John Benjamins. Amsterdam.

[30] Monachesi, Paola (2005): *The Verbal Complex in Romance. A Case Study in Grammatical Interfaces.* Oxford University Press, Oxford Studies in Theoretical Linguistics.

[31] Orasan C. (2000): A hybrid method for clause splitting in unrestricted English texts Available at: http://www.wlv.ac.uk/sles/compling/papers/orasan-00.pdf

[32] Passonneau, Rebecca; Diane Litman (1997): *Intention-based segmentation: human reliability and correlation with linguistic cues*, in Proc. 31th Annual Meeting of ACL, Ohio, pp. 148–155.

[33] Gheorghe Păun: *Marcus Contextual Grammars.* Kluwer Academic Publishers, Dordrecht, 1997.

[34] Pollard, Carl; Ivan Sag (1994): *Head-Driven Phrase Structure Grammar.* The University of Chicago Press, Chicago & London.

[35] Popârda, O.; N. Curteanu (2002): *L'évolution du discours juridique français analysé par la stratégie linguistique SCD.* In the volume "Representations du Sens Linguistique", Lagorgette, P. Larrivée (Eds.), LINCOM Europa, series Studies in Theoretical Linguistics, München, Germany, pp. 487–502.

[36] Puşcaşu, G. (2003): *Elementary discourse unit segmentation.* Dissertation thesis. "Al.I.Cuza" University of Iasi.

[37] Sgall, Petr; E. Hajičova, J. Panevova (1986): *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Kluwer Academic Publishers, Dordrecht.

[38] Soricut, R. and Daniel Marcu (2003): Sentence Level Discourse Parsing using Syntactic and Lexical Information. *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, May 27-June 1, Edmonton, Canada.

N. Curteanu,

Received February 21, 2006

Institute for Computer Science,
Romanian Academy, Iaşi Branch
B-dul Carol I, nr. 22A, 6600 IAŞI, ROMÂNIA
E–mail: *ncurteanu@yahoo.com* and *curteanu@iit.tuiasi.ro*