Published in: A.P. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications*, 1998, Oxford: Clarendon Press, 23-53.

Collocations and Lexical Functions

Igor A. Mel'čuk, University of Montreal

This paper discusses COLLOCATIONS from the viewpoint of their theoretical and practical (= lexicographic) description. Although they are, and have long been, a popular topic in linguistics, there is, as far as I know, no universally accepted formal definition of collocations nor a proposal for their uniform and systematic treatment. I hope to fill both these gaps, taking up the following four topics:

- Characterization and definition of collocations.
- Characterization and definition of Lexical Functions, the main tool for the description of collocations.
- Possible uses of Lexical Functions in linguistics.
- Presentation of Lexical Functions in the dictionary.

The literature on collocations is simply overwhelming. Since it is out of the question to present here even a partial survey thereof, I will abstain from references to other people's approaches, limiting myself to an absolute minimum.

1 Collocations

Collocations—no matter how one understands them—are a subclass of what is known as *set phrases* [1]; therefore, they have to be defined in terms of their *differentiae specificae* with respect to set phrases that are not collocations. This establishes my course: first, I define set phrases; then I propose a calculus or typology of set phrases; finally, I point out the place that collocations occupy among set phrases by supplying a formal definition of collocation.

1.1 Set phrases, or phrasemes

PEOPLE SPEAK IN SET PHRASES — rather than in separate words; hence the crucial importance of set phrases. At the same time, set phrases, or ***phrasemes***, represent one of the major difficulties in theoretical linguistics as well as in dictionary making.[2] Therefore, both linguistic theory and lexicography should really concentrate on them (this idea has been advocated for about 25 years by a number of people, such as, among others, Becker 1975, A. Pawley, R. Jackendoff, and the present writer). To show what I mean by phrasemes, here are several examples collected from one newspaper column (phrasemes are printed in bold):

(1) **a**. ***Of course***, *investors* ***accept the challenge offered*** *by this region*.
b. *Rabin* ***made*** *these* ***remarks*** *in an interview*.

c. *North Queensland is* ***best known*** *for its reef*.

d. *The rejection by the Bosnian Serbs of this plan* ***placed*** *them* ***on a collision course*** *with the five powers*.

e. *His statement* ***added fuel to the fire***.

f. *The share* ***price*** *of Perilya Mines* ***collapsed*** *yesterday* ***under the weight*** *of* ***heavy*** *selling* ***orders***.

g. *The* ***hardest thing***, ***for instance***, *will be* ***making decisions***.

A good dictionary of language L should include ALL phrasemes of L, because the main substantive property of a phraseme is its NON-COMPOSITIONALITY: it cannot be constructed, for a given Conceptual Representation, from words or simpler phrases according to GENERAL RULES of L; it has to be stored and used as a whole. A phraseme is a lexical unit; and, which is crucial, it is the numerically predominant lexical unit: in any language, phrasemes outnumber words roughly 10 to 1. Collocations make up the lion's share of the phraseme inventory, and thus they deserve our special attention.

1.2 Typology of phrasemes

1.2.1 Preliminary notions

To define a phraseme, I need some preliminary notions: those related to my assumptions concerning the way a speaker produces a text, those related to the linguistic sign, and two auxiliary concepts.

Text production

Let it be emphasized that the following discussion of collocations makes sense only if we look at them from the SPEAKER'S VIEWPOINT: in this paper, phrasemes are considered exclusively in terms of their production/construction (rather than in terms of their interpretation by the addressee).

I adopt the following view of the text production (= the Meaning-Text framework: Mel'čuk 1974, 9ff.; 1981; 1988a, 43-101; 1993, 41ff.):

— The speaker begins with what I call the Concept(ual) R(epresentation) [= ConceptR] of the situation he wants to verbalize. The ConceptR is a mental reflection of perceived reality, of the speaker's encyclopedic knowledge relevant to the situation in question, of his intentions, preferences, wishes and goals, of his ideas about the addressee, etc. The ConceptR of a given situation contains everything that might be needed in order to say what the speaker wants to say about it.

— Based on the initial ConceptR, the speaker constructs the Sem(antic) R(epresentation) [= SemR] of his intended utterance. He does it according to the Concepts-Meaning Model of his

language **L** [= CMM(**L**)], which associates with elements and configurations of the ConceptR elements and configurations of the corresponding SemR.

— From a given SemR, the speaker constructs, through a series of steps, the Phon(etic) R(e-presentation) of the utterance; he does it according to the Meaning-Text Model MTM(**L**), which associates with elements and configurations of the SemR all actual linguistic elements that make up the corresponding actual utterance.

An utterance is thus produced in two major steps using two models and involving three major representations:

$$\{\text{ConceptR}_k\} \overset{\text{CMM}}{\in} \{\text{SemR}_i\} \overset{\text{MTM}}{\in} \{\text{PhonR}_j\}$$

Linguistic sign

A phraseme (as well as a wordform, or a morph, etc.) is a ***linguistic sign***.

A linguistic sign is an ordered triple

$$\mathbf{X} = \langle \text{‘X’} ; /X/ ; S_x \rangle ,$$

where ‘X’ is the signified of the sign **X** (= its meaning), /X/ is its signifier (= its phonetic form), and S_x, its syntactics (= the set of data on its cooccurrence with other signs). See Mel’čuk 1982, 40-41 and 1993, 123ff. Except for syntactics, which has been added by the author, the concept of sign is clearly Saussurean.

Note that in the discussion of phrasemes, I leave syntactics out of consideration (for the sake of simplicity).

Auxiliary concepts

The concepts ‘unrestrictedly’ and ‘regularly’ (*constructed **E***), as applied to the signified or the signifier of a multi-unit expression are crucial to the definition of phraseme. These concepts are to be understood as follows:

1) *Unrestrictedly constructed **E*** = ‘**E** whose components are selected — for a given starting representation — according to **arbitrarily chosen** selection (≈ lexicon) rules of **L**’.

If the signified/the signifier **E** of an expression is constructed unrestrictedly, no rules $\{R_E\}$ applied to construct **E** are mandatory: instead of $\{R_E\}$, the speaker can apply ANY other applicable rules $\{R_{E'}\}$ to produce an equivalent **E’**. Thus, the signified and the signifier of the phrase *No parking* are not unrestrictedly constructed, because you are not supposed to express — on a sign — any equivalent meaning, for instance, ‘you should not park here’, or the same meaning in a different form — such as *Parking prohibited* or *Do not park*, although lexical (and grammatical) rules of English allow you to do so. On the contrary, the signified and the signifier of the sentence *This dictionary has been compiled by many people* are unrestrictedly

constructed, because you can express the same or an equivalent meaning by any other appropriate linguistic means: e.g., *This dictionary is the result of work by many hands*, etc.

ʻUnrestrictednessʼ thus means unlimited freedom of choice among (quasi-)equivalent independent meanings and expressions; it has to do with SELECTION of meanings and lexical units and is related to the concept 'selection rules of a language'.

Let me emphasize that for signifiers an additional proviso is necessary: A complex signifier is not unrestrictedly constructed if one of its parts is selected contingent on another one. We will see the importance of this condition in Definition 2.

2) *Regularly constructed* ***E*** = ʻ**E** whose components are combined **exclusively** according to general combination (♠ grammar) rules of **L**ʼ.

If the signified or the signifier **E** of an expression is constructed regularly, its components are put together, or united, solely by some general rules of **L**. Thus, all the expressions mentioned in the previous paragraph are constructed regularly, while the signified of the expression *the chip on* N*'s shoulder* ʻN's readiness to get angry and pick a fightʼ is not, because there is no way to construct it — out of the signifieds ʻchipʼ, ʻonʼ and ʻshoulderʼ — by general rules of English.

ʻRegularityʼ thus means observance of general rules in COMBINATION of meanings and expressions and is related to the concept of 'combination rules of a language'. These rules are represented in the formalism of the Meaning-Text theory by the *Operation of Linguistic Union* ⊕**:** putting together linguistic items of **L** while constructing expressions of higher order (Mel'čuk 1982, 41-42, and 1993, 139-144). The symbol ⊕ is reminiscent of arithmetical summation, but linguistic union is much more complex than simple addition: it presupposes observing ALL general combination rules of **L**, and this, in conformity with the nature of items being united (signified are united in a different way from signifiers and syntactics, etc.). Thus, **X**⊕**Y** denotes the regular union of signs **X** and **Y** (= the expression **X**⊕**Y** is regularly constructed out of signs **X** and **Y**); ʻXʼ⊕ʻYʼ is the regular union of signifieds ʻXʼ and ʻYʼ; etc.

Informally and approximately, a phraseme is a phrase whose signified and signifier CANNOT be constructed both unrestrictedly and regularly.

1.2.2 Free phrases

Definition 1: Free Phrase

A FREE phrase **A**⊕**B** in language **L** is a phrase composed of lexemes **A** and **B** and satisfying simultaneously the two following conditions:

1. its signified ʻXʼ = ʻA⊕Bʼ is unrestrictedly and regularly constructed on the basis of the given ConceptR — out of the signifieds ʻAʼ and ʻBʼ of the lexemes **A** and **B** of **L**;

2. its signifier /X/ = /A⊕B/ is unrestrictedly and regularly constructed on the basis of the SemR ‘A⊕B’ — out of the signifiers /A/ and /B/ of the lexemes **A** and **B**.

In prose, a free phrase **A⊕B** is a phrase such that: 1) its signified ‘A⊕B’ is freely constructed for the given ConceptR and can be replaced by any other sufficiently close signified ‘Y’, obtainable from the same ConceptR by rules of L; 2) this signified is a regular union of the signifieds of the phrase’s components and its signifier is a regular union of their signifiers, such that the phrase **A⊕B** can be produced according to general combination rules of L:

$$\mathbf{A}\langle\text{‘A’}; /A/\rangle \oplus \mathbf{B}\langle\text{‘B’}; /B/\rangle = \mathbf{A}\oplus\mathbf{B}\langle\text{‘A⊕B’}; /A\oplus B/\rangle$$

For a phrase to be free means FREEDOM OF SELECTION (of its signified and its signifier — with respect to the given ConceptR, that is, in the ultimate analysis, to the given situation; and of its signifier — with respect to the corresponding SemR) and FREEDOM OF COMBINATION (of its components: according to their own signifieds and syntactics plus general rules of L).

1.2.3 Set phrases = phrasemes

A SET phrase, or PHRASEME, **AB** is a phrase which is **not free**. Being not free can have three sources: both Conditions 1 and 2 in Def. 1 are violated; Condition 1 (but not 2) is violated; Condition 2 (but not 1) is violated:

1. Condition 1 is violated — such that the signified ‘X’ = ‘A⊕B’ is **not** unrestrictedly constructed on the basis of the given ConceptR (although it is regularly constructed) — and Condition 2 is violated as well (in the same way). Then for the given ConceptR, ONLY the given signified ‘A⊕B’ coupled with the given signifier /A⊕B/ is possible: the phrase in question is not unrestrictedly constructed. Not ALL applicable rules of L can actually be applied in the construction of **AB** whilw selecting its components; the choice of an appropriate meaning is reduced to one possibility (or to a few), and so is the choice of the form. As a result, we get PRAGMATIC PHRASEMES, or ***pragmatemes***. For instance, one sees on a restaurant sign *Caesar Salad: All you can eat;* its counterpart in French sounds *Salade César à volonté*, lit. ‘Caesar Salad to [your] wish = as much as you want’. It would be semantically and syntactically correct to say in French #*Salade César : Tout ce que vous pouvez manger*; however, this expression smacks of a calque: this is not the way the Frenchmen say it.[3] (The symbol # indicates pragmatic inappropriateness: #*X* means ‘*X* should not be used in the given situation’.) Thus kX: ALL YOU CAN EATl and kX À VOLONTÉl are pragmatemes of English and French, respectively.

2. Condition 1, but not Condition 2, is violated (as above). Then for the given ConceptR, still ONLY one given signified ‘A⊕B’ is possible, but it is unrestrictedly expressible, i.e., although you cannot use an equivalent meaning, for ‘A⊕B’ you can choose any one of several

possible (quasi-)synonymous expressions that rules of L allow. Such expressions are ***pragmatemes*** as well; for example, signs in an US library meant to prohibit talking say *No talking please*, *Please do not talk*, *Please be quiet*, etc. (but not, e.g., #*Don't make any noise please* or #*Keep silent please*).

All ready-made expressions (like greetings, typical phrases used in letters, conversational formulas, technical clichés, proverbs, sayings, etc.), even if semantically and syntactically they are 100 percent compositional, are pragmatemes: they are non-compositional pragmatically. (In this study, I will not consider pragmatemes: I am concerned solely with collocations.)

3. Condition 2 is violated, but Condition 1 is not (in the sense that the signified of **AB** is constructed unrestrictedly; yet it is not constructed regularly). Then for the given ConceptR, ANY signified obtainable by general selection rules is possible, but for a selected signified ‘A⊕B’, the corresponding signifier /X/ is not unrestrictedly constructed: if ‘X’ = ‘A⊕B’, then /X/ ≠ /A⊕B/. This gives us ***semantic phrasemes***.[4] (The important distinction between pragmatic and semantic phrasemes was first established in explicit terms in Morgan 1978.)

From now on, I will be considering the semantic phrasemes only. Let me establish their major types.

The violation of Condition 2 of Def. 1 can happen only in the following three ways:

• **AB** = ⟨‘C’; /A⊕B/⟩ | ‘C’⫰‘A’ & ‘C’⫰‘B’

This formula describes ***full phrasemes***, or ***idioms*** (k*[to] shoot the breeze*l, k*[to] spill the beans*l, k*[to] pull* [N's] *leg*l, k*[to] trip the light fantastic*l, k*of course*l, k*[to] put up*l, k*red herring*l). Instead of the regular union ‘A⊕B’ of the signifieds ‘A’ and ‘B’, an idiom **AB** has a different signified ‘C’, including neither ‘A’ nor ‘B’.

• **AB** = ⟨‘A⊕C’; /A⊕B/⟩ | ‘C’ is expressed by **B** such that /A⊕B/ is not constructed unrestrictedly

These are ***semi-phrasemes***, or ***collocations*** (*[to] land a* JOB; *high* WINDS; *[to] crack a* JOKE, *[to] do* [N] *a* FAVOR, *[to] give* [N] *an* ULTIMATUM, *[to] launch an* ATTACK, *[to] stand* COMPARISON [*with* N], *strong* COFFEE). The signified of a collocation includes ‘intact’ the signified of the one of its two constituent lexemes — say, of **A** (shown in the examples in small caps); **A** is freely chosen by the speaker strictly because of its signified. But the other component of its signified, i.e. ‘C’, is ‘problematic’: it is expressed by **B**, which is chosen CONTINGENT on **A** (this makes the signifier of a collocation to be not unrestrictedly constructed).

• **AB** = ⟨‘A⊕B⊕C’; /A⊕B/⟩ | ‘C’ ≠ ‘A’ & ‘C’ ≠ ‘B’

These are ***quasi-phrasemes***, or ***quasi-idioms*** (k*[to] give the breast*l [*to* N], k*[to] start a family*l, k*bacon and eggs*l, k*shopping center*l). Here the signified of **AB** includes the signifieds of both constituent lexemes, but contains as well an unpredictable addition ‘C’.

As a result, phrasemes can be classified as follows:

Phrasemes

Pragmatic Phrasemes		Semantic Phrasemes	
1. Pragmatemes	2. Idioms	3. Collocations	4. Quasi-idioms

1.3 The concept of collocation

Definition 2: Collocation (= Semi-Phraseme)

A COLLOCATION **AB** of L is a semantic phraseme of L such that its signified ‘X’ is constructed out of the signified of the one of its two constituent lexemes — say, of **A** — and a signified ‘C’ [‘X’ = ‘A⊕C’] such that the lexeme **B** expresses ‘C’ contingent on **A**.

The formulation “**B** expresses ‘C’ contingent on **A**” covers four major cases, which correspond to the following four major types of collocations:

1. **either** ‘C’ ≠ ‘B’, i.e., **B** does not have (in the dictionary) the corresponding signified;
   and [**a**. ‘C’ is empty, that is, the lexeme **B** is, so to speak, a semi-auxiliary selected by **A** to support it in a particular syntactic configuration;
   or **b**. ‘C’ is not empty but the lexeme **B** expresses ‘C’ only in combination with **A** (or with a few other similar lexemes)**]**;
2. **or** ‘C’ = ‘B’, i.e., **B** has (in the dictionary) the corresponding signified;
   and [**a**. ‘B’ cannot be expressed with **A** by any otherwise possible synonym of **B**;
   or **b**. ‘B’ includes (an important part of) the signified ‘A’, that is, it is utterly specific, and thus **B** is ‘bound’ by **A]**.

Examples (lexeme **A** is in small caps)

Case **1a**: collocations with support (= ‘light’) verbs, such as *[to] do a FAVOR*, *[to] give a LOOK*, *[to] take a STEP*, *[to] launch an APPEAL*, *[to] lay SIEGE* [*to* N].

Case **1b**: collocations such as *black COFFEE, French WINDOW*, Fr. *BIÈRE bien frappée* ⟨**battue*⟩ ‘well cooled [lit. ‘beaten’] beer’.

Case **2a**: collocations with intensifiers, such as *strong* ⟨**powerful*⟩ *COFFEE, heavy* ⟨**weighty*⟩ *SMOKER, deeply* ⟨**profoundly*⟩ *MOVED, [to] ILLUSTRATE vividly*.

Case **2b**: collocations such as *The HORSE neighs*, *aquiline NOSE, rancid BUTTER* or *artesian WELL*.[5]

Collocations constitute the absolute majority of phrasemes and represent the main challenge for any theory of phraseology. In order to describe collocations in a rigorous, systematic and exhaustive way, the Meaning-Text theory proposes the apparatus of Lexical Functions.

2 Lexical Functions

2.1 Introductory remarks

I begin with the general concept of ***Lexical Function*** [= LF] and then proceed to a particular one — ***Simple Standard*** LF, which is of special interest here (Zolkovskij and Mel'čuk 1967, Mel'čuk 1982, Mel'čuk *et al*. 1984, 1988, 1992).

The term *function* is used in its mathematical sense: **f**(x) = y, and the adjective *lexical* indicates that **f**'s domain of definition as well as the range of **f**'s values are both sets of lexical expressions.

A ***Lexical Function*** **f** is a function that associates with a given lexical unit [= LU] L, which is the ***argument***, or ***keyword***, of **f**, a set $\{L_i\}$ of (more or less) synonymous lexical expressions — the ***value*** of **f** — that are selected contingent on L to manifest the meaning corresponding to **f**:

$$\mathbf{f}(L) = \{L_i\}.$$

To put it differently, an LF, particularly a Simple Standard LF, is a very general and abstract meaning (coupled with a D(eep-)Synt(actic) role) which can be lexically expressed in a large variety of ways depending on the lexical unit to which this meaning applies.

About 60 Simple Standard LFs have been recognized so far in natural languages. Let me cite four preliminary examples and then proceed to definitions:

‹the one who/which undergoes ...› [*nomen patientis*]

$\mathbf{S_2}$(*to shoot*)	= *target*	$\mathbf{S_2}$(*to serve*)	= *client*
$\mathbf{S_2}$(*hotel*)	= *guest*	$\mathbf{S_2}$(*prison*)	= *prisoner*
$\mathbf{S_2}$(*doctor*)	= *patient*	$\mathbf{S_2}$(*hair-dresser*)	= *customer*

‹intense(ly)›, ‹very› [intensifier]

Magn($shave_N$)	= *close, clean*	**Magn**(*naked*)	= *stark*
Magn(*easy*)	= *as pie, as 1-2-3*	**Magn**(*thin*)	= *as a rake*
Magn(*to condemn*)	= *strongly*	**Magn**(*to rely*)	= *heavily*

‹do›, ‹perform› [support verb]

$\mathbf{Oper_1}(cry_N)$	= *to let out* [ART ~]
$\mathbf{Oper_1}(figure_N)$	= *to cut* [ART ~] [*He cut a miserable figure*]

$\mathbf{Oper_1}(strike_N)$ = *to be* [*on* ~]
$\mathbf{Oper_1}(support_N)$ = *to lend* [~]

‹realize›, ‹fulfill [the requirement of]›

$\mathbf{Real_2}(mine_N)$ = *to strike* [ART ~] [*Their car struck a land mine*]
$\mathbf{Real_2}(test_N)$ = *to withstand* [ART ~]
$\mathbf{Real_2}(joke_N)$ = *to get* [ART ~]
$\mathbf{Real_2}(exam_{N)}$ = *to pass* [ART ~]

[The symbol ART indicates that an article or a grammatically equivalent determiner should be used according to grammatical rules.]

2.2 Central concepts: LFs and Simple Standard LFs

Definition 3: Lexical Function

A function **f** associating with a lexical unit L a set **f**(L) of lexical expressions is called a ***Lexical Function*** if and only if one of the following two conditions is satisfied:

A. Either **f** is applicable to several Ls; in this case, for any two different L_1 and L_2, if $\mathbf{f}(L_1)$ and $\mathbf{f}(L_2)$ both exist, then:

1. Any elements of $\mathbf{f}(L_1)$ and of $\mathbf{f}(L_2)$ bear an (almost) identical relationship to L_1 and L_2, respectively, as far as their meaning and the DSynt-role are concerned; that is, for any $L_{\mathbf{f}(L_1)} \in \mathbf{f}(L_1)$ and any $L_{\mathbf{f}(L_2)} \in \mathbf{f}(L_2)$, it is true that

$$(L_{\mathbf{f}(L_1)}) : (L_1) \cap (L_{\mathbf{f}(L_2)}) : (L_2).$$

2. At least in some cases, $\mathbf{f}(L_1) \neq \mathbf{f}(L_2)$.

B. Or **f** is applicable to one L only (maybe to two or three semantically related Ls).

LFs of the type A are called ***normal*** LFs; those of the type B, ***degenerate*** LFs. In **f**(L), L is the ***keyword*** [6] of **f**, and **f**(L) is the ***value***.

Definition 4: Standard Lexical Function

A normal LF **f** is called a ***Standard Lexical Function*** if and only if the following two (additional) conditions are simultaneously met:

3. **f** is defined for a relatively large number of arguments. [To put it differently, the meaning (f) is sufficiently abstract and general to be applicable to many other meanings.]

4. **f** has a relatively large number of lexical expressions as its value — such that these expressions are more or less equitably distributed between different keywords.

Normal LFs that do not satisfy both Conditions 3 and 4, on the one hand, and degenerate LFs, on the other, are called ***Non-Standard***. (Thus the difference between Standard and Non-Standard LFs is purely quantitative: it concerns the number of possible keywords and value elements.)

Among Standard LFs, a subset of about 60 basic LFs is singled out: ***Simple*** Standard LFs; Simple Standard LFs constitute the foundations for the description of irregular derivation and restricted lexical cooccurrence (that is, of collocations).

2.3 The system of Simple Standard LFs

Since full lists of Simple Standard LFs are found in previously mentioned publications as well as in Mel'čuk and Zholkovsky 1988, Mel'čuk 1994, and then in Mel'čuk *et al*. 1984, 1988, 1992, I will not supply such a list here. Instead, I offer a brief substantive characterization of LFs, sketch their classification and quote four groups, which include the most current LFs.

2.3.1 Informal Characterization of LFs

Lexical Functions were first introduced by Zolkovskij and Mel'čuk (1967). They are used to describe two types of lexical phenomena that turn out to be of the same logical nature, that is, both are readily amenable to a description via the concept of function in the mathematical sense.

The first type involves ***paradigmatic*** lexical correlates $\{L_i'\}_{par}$ of a given LU L; they can be loosely described as (quasi-)synonymous with L. An L_{par}' can designate a situation or an object close to or identical with ʻLʼ, a generic notion for ʻLʼ, a situation implied by ʻLʼ, or a participant in the situation (implied by) ʻLʼ. Thus, for L = SCHOOL, $\{L_i'\}_{par}$ = TEACHER, STUDENT, SUBJECT, EXAM, LESSON, MARK, CLASS, [*to*] TEACH, [*to*] LEARN, ...; for L = [*to*] ESCAPE, $\{L_i'\}_{par}$ = [*to*] FLEE, k[to] BREAK AWAYl, $\text{ESCAPE}_{(\text{Noun})}$, ESCAPEE, kPLACE OF CONFINEMENTl, etc. Such lexical correlates show a kind of derivational relationships with L.

The second type involves ***syntagmatic*** lexical correlates $\{L_i'\}_{synt}$ of L that form with L collocations like some of those italicized in (1): *offer/accept the challenge, make a remark, best known, place on a course, heavy* [*selling*] *orders*. Thus, for L = SCHOOL, $\{L_i'\}_{synt}$ = TEACH (*school*), GO (*to school*), GRADUATE (*from a school*), ...; for $\text{L} = \text{ESCAPE}_{(\text{Noun})}$, $\{L_i'\}_{synt}$ = DARING.

LFs represent both types of lexical correlates of L.

2.3.2 Classification of LFs

LFs can be classified from different viewpoints; without having a scientific impact on the issue, such classifications facilitate the task of the user and thus possess pedagogical value. (The present classification and description of LFs follows some suggestions in Alonso and Tutin 1994.)

- Paradigmatic *vs*. syntagmatic LFs have been already characterized. Paradigmatic LFs deal with SELECTION; they are aimed at answering questions of the type "What do you call an object ⟨a situation⟩ X, related to Y?" — while speaking of X rather than of Y. Syntagmatic LFs deal with COMBINATION; they are aimed at answering questions of the type "What do you call the action ⟨characteristics, attribute, etc.⟩ X of Y?" — while speaking of Y rather than of X.

- Standard *vs*. non-standard LFs are different, first of all, with respect to the number of their possible keywords and value elements. Another important difference is that standard LFs participate in synonymic paraphrasing while non-standard ones do not (unfortunately, I cannot fully explain this difference here; see, e.g., Mel'čuk 1992b).

- 10 semantic/syntactic groups of Simple Standard LFs can be distinguished, based on the meaning and the DSynt-role associated with the given LF:

Basic LFs: **Syn**(onym), **Anti** [= antonym], and **Conv**(ersive)$_{ij}$. They embody the main semantic relations that play a special role in the MT-Theory — synonymy, negation, and converseness (*X precedes Y ~Y follows X*). Since they are relatively well known, I will not discuss them here, except to say that **Syn**, **Anti** and **Conv**$_{ij}$ can be semantically exact or approximate, i.e. they can have a richer ($_{\supset}$), poorer ($_{p}$), or intersecting ($_{i}$) meaning; in this case, they are quasi-synonyms, quasi-antonyms, and quasi-conversives. The same subscripts are also used for other LFs.

Derivatives are of two subtypes:

Syntactic derivatives represent nominalization **S$_0$** (*rejection* from REJECT), adjectivalization **A$_0$** (*urban* from CITY), verbalization **V$_0$** (*to attack* from [*the*] ATTACK), and adverbialization **Adv$_0$** (*well* from GOOD); **Pred** is a combination of a meaning with the copula; thus **PredMagn**(*animosity*) = *runs rampant*.

Semantic derivatives are, roughly speaking, agent noun **S$_1$**, patient noun **S$_2$**, active adjectival **A$_1$** (*in search* from [*to*] LOOK FOR), passive adjectival **A$_2$** (*under construction*

from [*to*] BUILD), place noun $\mathbf{S_{loc}}$, instrument noun $\mathbf{S_{instr}}$, active potential adjective $\mathbf{Able_1}$ (*inquisitive* from [*to*] ASK), passive potential adjective $\mathbf{Able_2}$ (*reliable* from [*to*] RELY), etc.

Generics: hyperonym **Gener** and metaphoric denotation **Figur** (*curtain of RAIN*).

Quantifiers: singulative **Sing** (*speck of DUST*), and collective **Mult** (*pride of LIONS*).

Modifiers: **Magn**, **Plus/Minus**, **Ver** (*restful SLEEP*), **Bon** (*valuable CONTRIBUTION, exquisite MEAL*).

Phasals: verbs denoting the three phases of an event — the beginning (**Incep**), the end (**Fin**), and the continuation (**Cont**). These LFs are often used combined with other verbal LFs.

Causatives: verbs denoting the three possible types of causation: causation of existence (**Caus**), causation of non-existence (**Liqu**), and non-causation of non-existence (**Perm**).

Note that the phasals stand in antonymous relation to each other; the same holds true of causatives: **Incep** = **AntiFin**, **Liqu = AntiCaus**, etc. Furthermore, causatives and phasals are also related, because you can cause the beginning, the end or the continuation of an event; however, I cannot go further into this problem.

Auxiliaries (= support, or light, verbs): semantically empty verbs linking a DSynt-actant [= A] of L to L; $\mathbf{Oper_{1,2}}$ takes L as its DSyntA **II**, $\mathbf{Func_{0,1,2}}$, as its DSyntA **I**, and $\mathbf{Labor_{12,21}}$, as its DSyntA **III** (for more details, see the following subsection).

Realizations: $\mathbf{Real_{1,2}}$, $\mathbf{Fact_{0,1,2}}$, $\mathbf{Labreal_{12,21}}$.

Varia: **Involv**, **Son**, **Imper**, **Degrad**, **Manif**, **Sympt**.

Simple Standard LFs can form combinations, to produce Complex Standard LFs: such as **Anti-Magn**, $\mathbf{IncepOper_{1,2}}$, $\mathbf{CausFunc_0}$, **CausPredPlus**, etc.

2.3.3 Illustrative list of LFs

As promised, I will quote in full four groups of Simple Standard LFs.

1. Semantic derivatives: actantial and circumstantial nouns

$\mathbf{S_i}$ is a standard name of the *i*-th DSyntA of L; it is thus an ACTANTIAL noun: in the first place, the agent [‘the one who L-s’] and the patient [= ‘the one whom people L’] nouns.

Syntactically, there are $\mathbf{S_i}$s of two types. An $\mathbf{S_i}$(L) of the first type is used in the text, as a rule, instead of L, especially if this L is a verb. An $\mathbf{S_i}$(L) of the second type is used in the text together with L, taking it as its own DSyntA $\overset{\mathbf{II}}{\cdots}$: $\mathbf{S_1}\overset{\mathbf{II}}{\cdots}$ L, etc.

$\mathbf{S_1}$(*to teach*) = *teacher*
$\mathbf{S_2}$(*to teach*) = (*subject*) *matter* , *subject* [in high school]
$\mathbf{S_3}$(*to teach*) = *pupil*

$\mathbf{S_1}$(*letter*) = *author*; *sender* [*of the letter*]
$\mathbf{S_2}$(*letter*) = *addressee* [*of the letter*]
$\mathbf{S_3}$(*letter*) = *contents* [*of the letter*]

$\mathbf{S_{instr}}$, $\mathbf{S_{med}}$, $\mathbf{S_{mod}}$, $\mathbf{S_{loc}}$, $\mathbf{S_{res}}$ — standard name of instrument, means, mode, location, and result of the situation denoted by L (as a rule, L is a noun or a verb); $\mathbf{S_n}$s are thus CIRCUMSTANTIAL nouns. Like actantial nouns, $\mathbf{S_n}$s normally are used instead of their keyword L; if they do not, they also take it as their DSyntA $\overset{\mathbf{II}}{\cdots}$: $\mathbf{S_{instr}}\overset{\mathbf{II}}{\cdots}\varnothing$L, etc.

$\mathbf{S_{instr}}$(*to shoot*) = *firearm*
$\mathbf{S_{instr}}$(*murder*$_{V,N}$) = [*murder*] *weapon*
$\mathbf{S_{med}}$(*to shoot*) = *ammunition*
$\mathbf{S_{loc}}$(*to fight* [as of two armies]) = *battlefield*
$\mathbf{S_{loc}}$(*war*) = *theater (of war)*
$\mathbf{S_{res}}$(*to learn*) = *knowledge, skills*
$\mathbf{S_{mod}}$(*to consider*) = *approach* [*I consider this problem as follows: ...~ My approach to this problem is as follows: ...*]

2. Intensifiers

Magn(*agree*) = *wholeheartedly*
Magn(*analysis*) = *trenchant*
Magn(*bore*$_N$) = *crashing*
Magn(*committed*) = *deeply*
Magn(*deserve*) = *richly*
Magn(*work*$_V$) = *as a Trojan, one's guts out*

3. Semi-auxiliary verbs

The LFs $\mathbf{Oper_i}$, $\mathbf{Func_i}$ and $\mathbf{Labor_{ij}}$ are ***support*** (or light)***verbs*** (cf. Gross 1981, Catell 1984); they are semantically empty (or emptied) in the context of the keyword LU. This LU is necessarily a noun whose meaning is or includes a predicate (in the logical sense of the term), thus presupposing actants. In other terms, the keyword of these LFs is, as a general rule, the name of an action, an activity, a state, a property, a relation, etc. (It can also be the name of a concrete object, which is defined by its role in a situation. Such is the case, e.g., of a body part or an organ: they represent what is called 'inalienable possession' and have as the value of the LF $\mathbf{Oper_1}$ the verb [*to*] HAVE or its equivalent.)

The support verbs serve to link, on the DSynt-level, (the name of) a DSynt-actant of L to L itself; they thus play an important semantic-syntactic role and can be loosely called semi-auxiliaries.

Oper$_i$ [Lat. *operari* ‘[to] do, carry out’]: the DSyntA **I** of this verb (and its SSynt-subject) is the phrase that is described in the Government Pattern [= GP] of L as the *i*-th DSyntA of L, and **Oper$_i$**'s DSyntA **II** (= its main [7] S(urface)Synt-object) is L itself. (Further DSyntAs of **Oper$_i$**, if any, are the phrases described in the GP of L as further DSyntAs of L.)

Oper$_1$(*blow$_N$*) = [*to*] *deal* [ART ~ *to* N] **Oper$_2$**(*blow$_N$*) = [*to*] *receive* [ART ~ *from* N]
Oper$_1$(*support$_N$*) = [*to*] *lend* [~ *to* N] **Oper$_2$**(*support$_N$*) = [*to*] *receive* [~ *from* N]
Oper$_1$(*order$_N$*) = [*to*] *give* [ART ~ *to* N] **Oper$_3$**(*order$_N$*) = [*to*] *receive* [ART ~ *from* N]
Oper$_1$(*resistance*) = [*to*] *put up* [ART ~], **Oper$_2$**(*resistance*) = [*to*] *meet* [ART ~],
[*to*] *offer* [ART/0 ~] [*to*] *run* [*into* ART ~]
Oper$_1$(*control$_N$*) = [*to*] *have* [~ *over* N] **Oper$_2$**(*control$_N$*) = [*to*] *be* [*under* N's ~]

The expression in brackets following each element of the value of the LF illustrated is its ***reduced Government Pattern*** — its lexical subentry.

Func$_i$ [Lat. **functionare* ‘[to] function’]: the DSyntA **I** of this verb (and its SSynt-subject) is L itself, and its DSyntA **II** (= its main SSynt-object) is the *i*-th DSyntA of L.

Func$_1$(*blow$_N$*) = *comes* [*from* N] **Func$_2$**(*blow$_N$*) = *falls* [*upon* N]
Func$_1$(*proposal*) = *comes, stems* [*from* N] **Func$_2$**(*proposal*) = *concerns* [N]

In cases where there is no object at all, i.e. **Func$_i$** is an absolutely intransitive verb, the subscript **$_0$** is used:

Func$_0$(*snow$_N$*) = *falls* [*At night, the snow started to fall*]
Func$_0$(*war*) = *is on*
Func$_0$(*silence*) = *reigns*

Labor$_{ijk}$ [Lat. *laborare* ‘[to] work, toil’]: the DSyntA **I** of this verb (and its SSynt-subject) is the *i*-th DSynt-actant of L, its DSyntA **II** (= its main SSynt-object) is *j*-th DSyntA of L, its DSyntA **III** (= its SSynt-object) is *k*-th DSyntA of L, and its further DSyntA (= its third SSynt-object) is L itself.

Labor$_{12}$(*interrogation*) = [*to*] *subject* [N *to an interrogation,* where the keyword INTERROGATION is DSyntA **III** of the verb *subject*]

Labor$_{321}$(*lease*$_N$) = [*to*] *grant* [N *to* N *on lease*, where the keyword LEASE$_N$ is DSyntA **IV** of the verb *grant*]

Oper$_{0/i}$, **Func**$_{0/i}$ and **Labor**$_{ijk}$ can be paired in converse relations:

$$\mathbf{Oper}_1 = \mathbf{Conv}_{21}(\mathbf{Func}_1);\ \mathbf{Labor}_{12} = \mathbf{Conv}_{132}(\mathbf{Oper}_1);\ \text{etc.}$$

These relations may be diagrammed — for a two-actant LU — as follows:

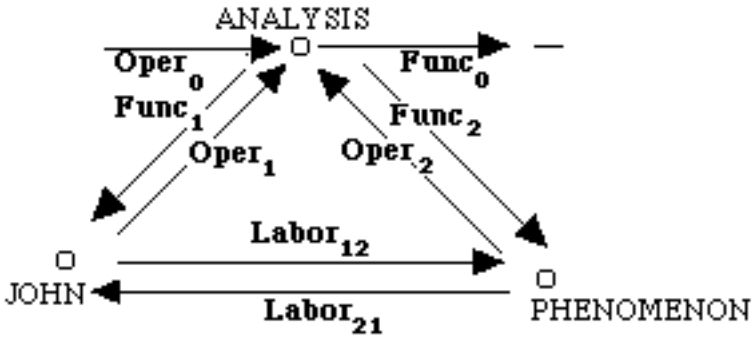


Fig. 1 Support verbs and their DSynt-relationships with their keyword

In Fig. 1, a two-actant lexeme L (= ANALYSIS, with two DSyntAs: **I** — JOHN, and **II** — PHENOMENON) is presented; the whole means ‘John analyzes the phenomenon’. The arrows represent the LFs values, i.e. the support verbs in question; the arrow’s tail indicates DSyntA **I** of the support verb (= Grammatical Subject), the head pointing to its DSyntA **II** (= Main Object). Thus:

Oper$_1$(*analysis*) = [*to*] *carry out* [*John carries out the analysis of the phenomenon*];

Oper$_2$(*analysis*) = [*to*] *undergo* [*The phenomenon underwent (careful) analysis (by John)*];

Func$_1$(*analysis*) = *is due* [*The analysis of this phenomenon is due to John*];

Func$_2$(*analysis*) = *covers*, *concerns* [*John’s analysis concerns this phenomenon*];

Labor$_{12}$(*analysis*) = [*to*] *submit* [*John submits this phenomenon to a (careful) analysis*];

Labor$_{21}$(*analysis*) = — (it could be something like **The phenomenon prompts John to an analysis*);

Func$_0$(*analysis*) = *takes place*, *occurs* [*John’s analysis of the phenomenon is taking place*];

Oper$_0$(*analysis*) = [*one*] *sees* [*One sees an analysis of the phenomenon by John*].

A different way to express the same idea is by using the following table:

DSynt-role of L and its DSynt-actants with respect to the support verb V^{LF} Support verb V^{LF}	DSynt-actant **I** of V^{LF} is:	DSynt-actant **II** of V^{LF} is:	DSynt-actant **III/IV** of V^{LF} is:
Oper$_{1/2}$	**I**st /**II**nd DSyntA of L	L	————————
Func$_{0/1/2}$	L	none/**I**st / **II**nd DSyntA of L	————————
Labor$_{12/21}$	**I**st /**II**nd DSyntA of L	**II**nd / **I**st DSyntA of L	L

Fig. 2 Definitions of the support verbs

From the diagram of Fig. 1 and the table of Fig. 2 it is easy to see why the support verbs are presented as THREE LFs: these LFs are distinguished according to their syntactic behavior with respect to the major sentence SSynt-elements, and there are three such elements: Grammatical Subject, Main (roughly, Direct) Object and Second (roughly, Indirect or Prepositional) Object.

4. Realizations

Real$_{0/i}$, **Fact$_{0/i}$** and **Labreal$_{ij}$**, or ***fulfillment verbs***, mean, roughly, ‘[to] fulfill the requirement of L’ [= ‘[to] do with L what you are supposed to do with L’] or ‘L fulfills its requirement’. The "requirements" differ with respect to different Ls: thus the "requirement" of a hypothesis is its confirmation, and the "requirement" of a disease is the malfunctioning/death of the person affected, while the "requirement" of an artefact is that it be used according to its intended function. **Real$_i$** [Lat. *realis* ‘real’], **Fact$_{0/i}$** [Lat. *factum* ‘fait’] and **Labreal$_{ij}$** [a hybrid of **Labor** and **Real**] are (more or less) synonymous full verbs, different with respect to their syntax only; their keywords are nouns whose meaning includes the component corresponding to a "requirement": ‘supposed to ...’, ‘designed to ...’, etc.

In sharp contrast to the support verbs, which accept as their keywords basically abstract nouns, the fulfillment verbs can have both abstract and concrete keywords, provided the latter have actants and imply a "requirement". Such concrete nouns are necessarily the names of artefacts or organs, which are by definition ‘designed to ...’.

Syntactically, **Real$_i$**, **Fact$_{0/i}$** and **Labreal$_{ij}$** are fully analogous to the LFs **Oper$_i$**, **Func$_{0/i}$** and **Labor$_{ij}$**, respectively. This means that the keyword L and its DSyntAs fulfill with respect to **Real$_i$** the same syntactic roles as they do with respect to **Oper$_i$**, etc. Therefore, they are linked to their keywords in the following way:

II **I** **III**

Real$_{0/i}$——· L, **Fact**$_{0/i}$——· L, and **Labreal**$_{ij}$——· L.

Real$_1$(*accusation*) = [*to*] *prove* [ART ~] **Real**$_2$(*law*) = [*to*] *abide* [*by* ART ~]

Real$_1$(*car*) = [*to*] *drive* [ART ~] **Real**$_2$(*hint*) = [*to*] *take* [ART ~]

Real$_1$(*illness*) = [*to*] *succumb* [*to* ART ~] **Real**$_2$(*demand*) = [*to*] *meet* [ART ~]

Real$_1$(*bus*) = [*to*] *drive* [ART ~] **Real**$_2$(*bus*) = [*to*] *ride* [*on* ART ~]

Compare:

Oper$_1$(*obstacle*) = [*to*] *face* [*an* ~], but **Real**$_1$(*obstacle*) = // [*to*] *turn back*

Oper$_2$(*attack*) = [*to*] *be* [*under an* ~_.*of* N], but **Real**$_2$(*attack*) = [*to*] *fall* [*to* ART ~_.*of* N]

Oper$_2$(*exam*) = [*to*] *take* [ART ~], but **Real**$_2$(*exam*) = [*to*] *pass* [ART ~]

Fact$_0$(*hope*$_N$) = *comes true* **Fact**$_0$(*film*$_N$) = *is playing, is on*

ContFact$_0$(*luck*) = *holds* Rus. **ContFact**$_1$(*udača* ‹luck›) = *ne pokidaet* [N$_{acc}$],

Labreal$_{12}$(*gallows*) = [*to*] *string up* [N *on* ART ~] lit. ‹does not abandon›

Labreal$_{12}$(*saw*) = [*to*] *cut* [N *with* ART ~]

2.4 LFs and collocations

LFs cover ALL collocations with the only exception of those that are covered by the Government Pattern [= GP] of L: cf. Fr. *assurance vie* ‹life insurance›, where life is what you insure, vs. *assurance maladie*, lit. ‹illness insurance›, where illness is what you insure against (cf. *health insurance*); similarly, *assurance auto* ‹car insurance› vs. *assurance incendie*, lit. ‹fire insurance›; etc. The restricted cooccurrents in these collocations are Sem-actants of the keyword. Further examples include *un condamné à mort*, lit. ‹a person-sentenced to death›, vs. *un condamné à vie*, lit. ‹a person-sentenced to life [in prison]›; Fr. *auto-école* vs. Eng. *driving school*; or else *sick leave* ~ *maternity leave* ~ *study leave*; *hit list* ~ *shopping list*; *life sentence*, etc. All these collocations are described not by LFs of the keyword L, but by L's GP.

On the other hand, NOT ALL LFs describe collocations: only the syntagmatic LFs do (whereas the paradigmatic LFs represent the derivatives of the keyword). Thus the set of all collocations and that of all expressions described by LFs overlap: they have an important intersection.

2.5 The degree of fixedness of LF expressions

An important property of LF expressions (and of course of the collocations they represent) is ***fixedness***: the quantity of similar phrasemes that exist for the phraseme under consideration The phrase *pay attention* is very fixed: **Oper**$_1$ is expressed as PAY only with ATTENTION (even **pay heed* is impossible), VISIT/CALL and COMPLIMENT (**pay a*

greeting). On the other hand, the phrase *give [him] a look* is much less fixed: **Oper$_1$** is expressed as GIVE with scores of nouns (*give* [N] *a pull, a punch, a smile, a tug, a push, a kick, a stroke, a kiss, a try, a greeting*, etc.). Moreover, it is possible to characterize semantically the resulting construction: it means ‘[affect an object or communicate with a being] voluntarily, performing one unit of the activity involved’ (Dixon 1991, 348-51). With body parts, **Oper$_1$** is fully predictable: it is always HAVE. However, since in very many cases **Oper$_1$** is expressed by phraseologically bound LUs, all the expressions with **Oper$_1$** are considered phrasemes. The same is also true about all LF expressions: once LF, always LF. As a result, we can have LF expressions (i.e., collocations) with a very low degree of fixedness: the elements of the LF value may be semantically transparent and their cooccurrence, predictable — yet the expressions in question remain phrasemes by analogy with more restricted cases. The degree of fixedness is thus an independent parameter of phrasemes, cutting across their defining properties (restrictedness of selection and irregularity of combination).

As a result, in spite of the basically idiosyncratic character of LFs, in many cases a given LF has the same values for quite a few different keywords, the reason often being semantic proximity: semantically related LUs can feature the same values for a given LF. This fact can be accounted for by following the general principle of lexical inheritance (Mel’čuk and Wanner 1994):

Principle of Lexical Inheritance

> All lexicographic data shared by a family of semantically related LUs should be stored just once — under one LU of the corresponding vocable or under the generic LU of the corresponding semantic field, from where these data are "inherited" in each particular case.

This principle covers, first of all, LFs; however, I cannot explain here all the techniques of generalizing over common values of LFs.

2.6 LFs *vs.* semantic restrictions

Not all cases of restricted cooccurrence of LUs are cases of restricted LEXICAL cooccurrence. Thus consider the Russian verb OSIBIT´SJA + N_{instr}, roughly ‘use or try to use the wrong N’:

(2) **a**. *On osibsja dver´ju* ‘He passed or tried to pass through the wrong door’.
vs.
**On osibsja ključom* ‘He used or tried to use the wrong key’.

b. *On osibsja adresom* ‘He came to a wrong address’.
vs.
**On osibsja avtobusom* ‘He boarded or tried to board the wrong bus’.

c. *On osibsja nomerom* ‘He called a wrong (telephone) number’.

vs.

**On osibsja knigoj* ‘He took or tried to take a wrong book’.

The cooccurrence of OSIBIT´SJA in this construction (which, astonishingly, no Russian dictionary I have consulted stores) is extremely limited and looks very capricious. However, the expressions *osibit´sja dver´ju* ⟨ *adresom, nomerom*⟩ (and a few others that are possible) are not collocations, but free phrases: the restrictions observed are purely semantic. The meaning of the verb OSIBIT´SJA here is ‘mistakenly try to establish contact with somebody at a location identified by Y’ while believing that Y’ is Y’’’; therefore, Y can be only something that might be interpreted as being or identifying a location. Stretching things a little, one might say

d. *On osibsja čemodanom* ‘He opened a wrong suitcase’;

this sentence, however, cannot mean *‘He took/bought/brought the wrong suitcase’. As we see, one has to distinguish, on the one hand, between a very specific and therefore highly restrictive meaning — that is, SEMANTIC constraints in lexicographic definitions, and, on the other, genuine LEXICALLY restricted cooccurrence of LUs. Only the latter comes under the jurisdiction of LFs.

3 LFs in linguistic applications

To illustrate the role of LFs in linguistics, I will say a few words about their possible uses in the area known as Computational Linguistics. More specifically, I will touch upon LFs in Automatic Translation and Text Generation. Four aspects are of particular interest: lexical choices, paraphrasing, communicative structure, and text cohesion.

3.1 LFs and lexical choices (collocational aspect)

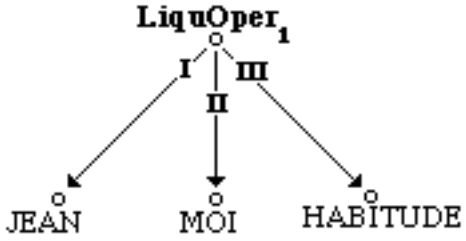
Suppose a system of Automatic Translation in which the transfer (from the source language into the target language) is done at the level of DSynt-Structure [= S]; suppose furthermore that we are interested in the translation of collocations. In such a case, it suffices to reduce the source language collocation to its LF-representation, then translate the keyword only and, finally, to select the value of the LF for the equivalent of the keyword in the target language. For instance, consider the French sentence (3):

(3) *Jean m’a détourné de cette habitude* ‘John broke me of this habit’.

- Analysis

Using a monolingual French dictionary which lists the values of all LFs for all head LUs (plus of course all syntactic mechanisms needed), (3) becomes (3’):

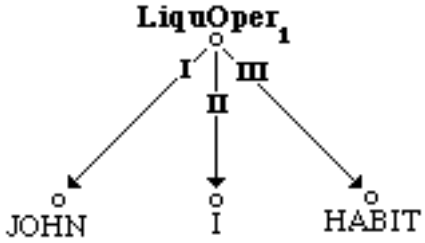
(3') = DSyntS(3)



- Transfer

Using a bilingual (or multilingual) network of lexical correspondences, the French tree (3') is replaced with the English tree (4'):

(4') = DSyntS(4)



- Synthesis

Using a monolingual English dictionary, again with the values of all LFs specified for all head LUs (and all corresponding syntactic mechanisms), the tree of (4') is turned into the English sentence (4):

(4) *John broke me of this habit*.

As can be gleaned from this simplified example, for collocations only the keywords need actual transfer, i.e. looking up of their equivalents (Fr. HABITUDE = Eng. HABIT). The search for the "bizarre" correspondence DÉTOURNER = [*to*] BREAK in the context of HABIT is avoided altogether: [*to*] BREAK [N *of* ~] is computed as an element of the value of the LF **LiquOper$_1$**(HABIT) in an English dictionary independently of the source language. LFs play thus the role of a transfer interlingua. In this way, multilingual translation does not require many pairwise-arranged transfer dictionaries of collocations. It is enough to have monolingual dictionaries with LFs specified plus indexes of multilingual translation equivalents for keywords only.

The same type of procedure can be used by any system of Text Generation that produces the output text passing by a DSynt-Structure. One such system is described in Iordanskaja, Kim and Polguère 1994: it makes active use of LFs for lexical choices and, in particular, does

so with an eye to paraphrasing and the Communicative Structure of the sentence to be generated (see **3.2** and **3.3**).

To make things clearer, let me cite a series of "bizarre" correspondences that can be easily and naturally expressed in terms of LFs:

	Eng. HABIT <> Fr. HABITUDE	
IncepOper$_1$	acquire, develop, form [ART ~], get [into ART ~], take [to ART ~]	contracter, prendre [ART ~]
FinOper$_1$	drop [ART ~], drop [ART ~], get out [of ART ~], get rid [of ART ~], ...	abandonner, perdre [ART ~]
LiquOper$_1$	break [N of ART ~], wean away [N from ART ~]	détacher, détourner [N de ART ~]

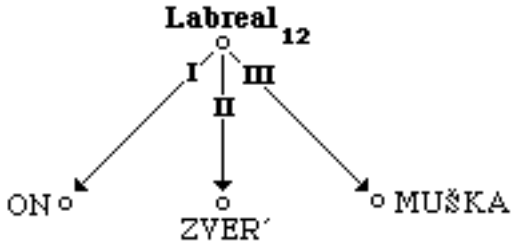
3.2 LFs and paraphrasing (syntactic aspect)

A well-known thorny problem of text generation is widespread incompatibility of a given LU and the syntactic construction in which it must appear; in many a case, a lexical choice made entails syntactic restructuring. LFs turn out to be helpful in this respect as well. The fact is that the equations relating LFs [9] allow for a number of important syntactic transformations. Thus, consider the Russian sentence (5), which has to be translated into English:

(5)*On vzjal zverja na musku*, lit. ‘He took the-beast on bead = took aim at the beast’.

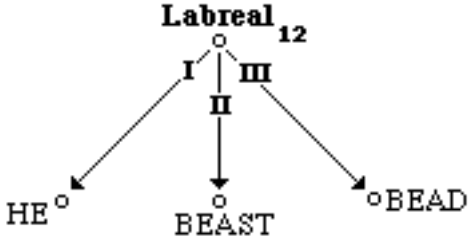
Under analysis, it is reduced to the DSyntS (5’):

(5’) = DSyntS(5)



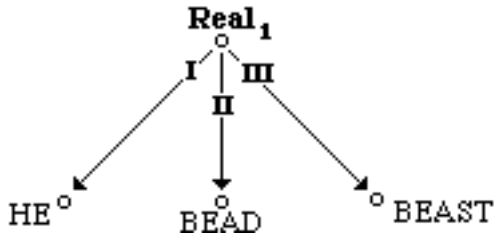
Under transfer, the Russian nominal lexemes are replaced with their equivalents: HE, BEAST, and BEAD. But the resulting English DSyntS (6’):

(6’)



cannot be implemented directly, because the English equivalent of MUSKA , i.e. BEAD, does not have a **Labreal**$_{12}$: **Labreal**$_{12}$(*bead*) = ? Yet BEAD has a **Real**$_1$: [*to*] DRAW. Replacing **Labreal**$_{12}$ by **Real**$_1$ and performing, at the same time, the standard transformation associated with **Labreal**$_{12}$ <=> **Real**$_1$ substitution, we obtain the correct English tree (6''):

(6'')



Under synthesis, this DSyntS is realized as (6):

(6) *He drew a bead on the beast*.

which is the optimal translation of (5). In this way, LFs take upon themselves the syntactic adjustments needed to carry out the transfer between languages — in cases where the LFs are involved.

Another telling illustration of the process is the translation of the English sentence (7) in Russian:

(7) *He was stabbed three times, once fatally*.

Its translation appears as (8):

(8) *Emu bylo naneseno tri nozevyx rany, odna iz kotoryx okazalas´ smertel´noj*, lit. ‘To-him was dealt three knife wounds, one of which turned-out mortal’.

Again, if a translation system tries to make the transfer at the level of DSyntS, it can use, to obtain the result shown, the following LFs (the description is fragmentary and approximate; it gives only a rough idea of how such transfers can occur):

Analysis (English)

[*to*] *stab* [N] = **Labreal**$_{12}$(*knife*) + **CausFunc**$_1$(*wound*)

<<*fatal* = **Magn**(*wound*) [the symbol "<<" indicates the extreme value of **Magn**]

Transfer

Eng. KNIFE = Rus. NOZ

Eng. WOUND = Rus. RANA

Synthesis (Russian)

CausFunc$_1$(*rana*) = *nanesti* [N_{dat} ~u_{acc}]

Magn(*rana*) = <<*smertel´naja*

caused with knife(*rana*) = *nozevaja*

If, however, a translation system proceeds via a SemR, then its task (in regard to restricted lexical cooccurrence) is to establish the relevant LF starting from the initial SemR

and then to "compute" its value for the given L, based on a monolingual dictionary of the ECD type. Of course the same procedure is needed for text generation, whatever its underlying representation.

3.3 LFs and the Communicative Structure of sentences (communicative aspect)

Suppose a text-generation system has to verbalize the meaning of sentence (9) (this example is adapted from Wanner and Bateman 1990, where the use of LFs in connection with the Communicative Structure of the sentence is discussed):

(9) *The adjective "electronic"* ***indicates*** *to the reader that the dictionaries are dedicated to computers.*

If in the Semantic Structure of (9) the meaning of the phrase *the adjective "electronic"* is specified as the theme, then (9) can be produced. But if the meaning of the phrase *to the reader* is specified as the theme, a different syntactic structure is needed, which will eventually lead to (9'):

(9') *The reader* ***gets an indication*** *that the dictionaries are dedicated to computers from the adjective "electronic".*

To replace *indicate* with *get an indication*, one needs the paraphrasing equations of the type

$$V \Longleftrightarrow \mathbf{S_0}(V) + \mathbf{Oper_2}(\mathbf{S_0}(V))$$

[*X analyses Y* $\Longleftrightarrow$ *Y undergoes an analysis by X*,
X resists Y $\Longleftrightarrow$ *Y runs into resistance from X*,
X orders Y to do Z $\Longleftrightarrow$ *Y receives from X an order to do Z*, etc.],

and most importantly, a dictionary which specifies, for each L, the values of LFs (cf. Mel'čuk 1992b).

3.4 LFs and text cohesion (cohesional aspect)

LFs prove equally useful in selecting the referring expressions in anaphorical links in such a way as to avoid tedious repetitions and guarantee, at the same time, the maximum cohesion of the resulting text (see Tutin 1992 and Alonso *et al*. 1992, 160-165). Thus, speaking of an *ambush*, you can refer back to it by calling its participants *attackers*:

(10) *An Indonesian* ***patrol*** *was caught in an* ***ambush***. *The* ***attackers*** *fired three rockets at the* ***soldiers*** *and sprayed them with automatic fire.*

Here, *attacker* = $\mathbf{S_1}$(*ambush*), and *soldier* = $\mathbf{S_1}$(*patrol*). This lexical knowledge is used to construct the sentence sequence (10) in an obvious way. [10]

Another example:

(11) *Sales **increased** slightly in Quebec and Ontario. Modest **gains** were also reported in British Columbia.*

Instead of simply repeating the same phrase and saying *Sales also increased slightly in British Columbia,* the speaker chooses to use $\mathbf{S_2}$(*increase*) = $gain_N$ [(the amount by which X increased)],which allows him to produce a more varied and elegant text.

4 LFs in the lexicon

LFs are specified — for each LU — in the dictionary, so that they are essentially a lexicographic problem. The MTT presupposes the existence of a special type of lexicon in whose entries LFs occupy an important place and which constitutes one of the central modules of the MT model of natural language. This lexicon is the **E**xplanatory **C**ombinatorial **D**ictionary [= ECD]. For a working understanding of LFs a brief description of their representation in an ECD is indispensable; this, in its turn, requires a cursory sketch of the ECD. Since, however, the publications on ECDs are numerous (Zolkovskij and Mel'čuk 1967, Mel'čuk *et al*. 1984, 1988, 1992, Mel'čuk and Zholkovsky 1984, 1988, Mel'čuk and Polguère 1987, Mel'čuk 1988b, 1989, 1992, Ilson and Mel'čuk 1989), I will limit myself to a very short characterization.

4.1 Main properties of an ECD

The ECD is semantics- and paraphrase-based: (quasi-)synonymous paraphrases constitute the main target as well as the main research tool for an ECD. Its entries are supposed to supply ALL lexical information which might be needed for the two tasks that any linguistic model has to tackle:

- the transition from a Semantic Representation (formally, a network composed of semantic units) to a DSynt-R(epresentation; formally, a dependency tree composed of actual LUs);
- the construction, for a given DSyntR, of all the DSyntRs which are (up to the communicative organization) synonymous with it; this is paraphrasing.

The main SUBSTANTIVE property of an ECD is that it is a PHRASAL DICTIONARY. It contains set phrases, i.e. phrasemes, 1) as headwords of numerous entries (idioms and quasi-idioms) and 2) as important data within the entries (semi-idioms, i.e. collocations, represented as LF-expressions).

The six main FORMAL properties of an ECD are as follows:

1) An ECD is is a theoretical dictionary: it is elaborated within a coherent linguistic theory, featuring developed semantic, syntactic and morphological components, or modules, and putting a strong emphasis on the lexicon.

2) An ECD is an active dictionary: it is consistently geared to production, or synthesis.

3) An ECD is a semantic dictionary: it is based on semantic representations of all the expressions it contains, the definition being the central part of a lexical entry.

4) An ECD is a combinatorial dictionary: it is centered around restricted cooccurrence (syntactic and lexical).

5) An ECD is a formalized dictionary: it can be considered as a lexical database.

6) An ECD tries to be exhaustive with respect to individual LUs (lexemes and phrasemes): a lexical entry includes whatever a native speaker knows about the LU in question.

4.2 The structure of an ECD article

All LUs stored in an ECD have dictionary articles of the same structure. An ECD article is divided into three major zones:

• The SEMANTIC zone: the definition (= a SemR of the head lexical unit L), which (in the case of LUs with predicative meaning) is based on a propositional form with variables for semantic actants and constitutes a strict decomposition of the meaning of L. For instance, the verb [*to*] HELP (in one of several senses):

X helps Y to Z with W = (Y trying to do or doing Z,‖ X uses X's resources W, adding W to Y's efforts with the goal that W facilitates for Y doing Z).

(The part to the left of "‖" symbol is a presupposition: it remains asserted when the entire meaning of HELP is negated: *John didn't help Mary to prepare the dinner* still implies that Mary prepared the dinner.)

The LFs of L are semantically related to some particular semantic components of L's definition. Thus, **Magn**(*help*) = *a lot* intensifies ‹facilitate›; the same is true for all LFs.

• The SYNTACTIC zone: the Government Pattern (= a subcategorization frame), which specifies, for each Sem-actant, the corresponding DSyntA and lists all surface means of expressing it in the text. Cf. the Government Pattern for the verb (*to*) HELP [C stands for 'column', so that $C_{III.1}$ means (column III, line 1)]:

	Y = II	Z = III	W = IV
1. N	1. N	1. V_{inf} 2. *to* V_{inf} 3. *with* N 4. *in* V_{ing} 5. $PREP_{dir}$ N	1. *with* N 2. *by* V_{ger}

1) $C_{III.1}$: ‘X being directly involved in Z’ [= ‘X doing Z himself’]

2) $C_{III.2}$: ‘X not being directly involved in Z’ [= ‘X not doing Z himself, but providing some external resources to Y’] [11]

3) $C_{III.5}$: **if** Z = ‘travel / move [something] in the direction α’, **then** [III = L(‘α’) **and** $C_{III} = C_{III.5}$] **is possible**

4) $C_{III.3} + C_{IV.1}$: **undesirable**

Impossible : **Kathleen was helped move the furniture (by Arthur, and not by Jane)* [correct expression: *...**to** move the furniture*] (General rule of English syntax: no bare infinitive with the passive).

Undesirable: ?*Kathleen helped Arthur **with** his work **with** her advice* [correct expression: either *... **in** his work **with** her advice* or *... **with** his work **by** advising him*] (Rule 4).

Kathleen helped the old gentleman finish his preparations ⟨ helped the boy to finish his studies with her generous financial assistance, helped me in buying my last car with her advice, helped Jack out of his coat, helped Jack up the stairs by a kick in the bottom /by giving him a firm push/shove⟩ .

LFs of L are related to L’s GP in an obvious way: thus, $\mathbf{Oper_1}$ is different from $\mathbf{Labor_1}$ in so far as the former takes L as its DSyntA **II** and the latter, as its DSyntA **III**; $\mathbf{AntiBon_1Involv}$(*car*) = *smash* [*into* N] (it is the DSyntA **I**, i.e. the car, that suffers), while $\mathbf{AntiBon_2Involv}$(*car*) = *run over* [N] (it is the DSyntA **II**, i.e. the person run over, that suffers); etc. Moreover, the values of many LFs have GPs of their own; however, I cannot develop this point here.

• The LEXICAL COOCCURRENCE zone: Lexical Functions, which present the RESTRICTED LEXICAL COOCCURRENCE of the headword L. The description of restricted lexical cooccurrence of L is fully adjusted to L's definition and to its Government Pattern.

4.3 A sample lexical entry, ECD-style

By quoting a full-fledged lexical entry I hope to show LFs in their natural habitat, that is, in a dictionary. Among other things, it can be seen how the LFs in the entry are related to the definition and to the Government Pattern.

REVULSION

X's revulsion for Y = X's (strong) negative emotion about Y similar to what people normally experience when they are in contact with something that makes them sick and such that it causes that X wants[12] to avoid any contact with Y.

Government Pattern

X = I	Y = II
1. N's 2. A_{poss}	1. *against* N 2. *at* N 3. *for* N 4. *toward* N

1) $C_{II.2}$: N denotes something that happens and and can be seen or felt
2) $C_{II.4}$: N denotes people

John's ⟨*his*⟩ *revulsion against racism* ⟨*against greed/dismal results of his endeavor*⟩; *John's* ⟨*his*⟩ *revulsion at such behavior* ⟨*at the sight of sea food*⟩; *John's* ⟨*his*⟩ *revulsion for work* ⟨*for all those killings*⟩; *John's* ⟨*his*⟩ *revulsion for* ⟨*toward*⟩ *these scoundrels / toward the government*; *John's* ⟨*his*⟩ *revulsion *at these shouts* [correct: *...for these shouts*]

Lexical Functions

Syn_p : distaste
Syn_i : repugnance; repulsion; disgust; loathing
$Anti_i$: attraction
$Conv_{21}Anti_i$: appeal
A_1 : revulsed
$Able_2$: revulsive

$Magn + Able_2$	: of utmost ~ | G = SCENE, SIGHT [G stands for the syntactic Governor of the LF value]
Magn	: deep < extreme << utmost
AntiMagn	: slight
Adv_1	: in [~]
Propt	: from [~]
$Oper_1$	: experience, feel [~]
$Magn + Oper_1$	: be filled [with ~]
$Magn + Labor_2$	: fill [N with ~]
$Conv_{21}Caus_2Oper_1$	: be driven [to ~]
Adv_1Manif	: with [~]

Examples

He did it from deep revulsion against the bitterness of the sectarian strife. Any revulsion they might feel from fat-ass bastards they ran up against professionally was *ad hominem* and not *ad genus* [A. Lurie]. Kathleen turned her head away in revulsion. I felt no revulsion for her maternal phantasies, only a practical concern. She met his advances with revulsion. It was a scene of utmost revulsion. Pam was driven to revulsion (by the sight of the dead animal) ⟨ *The sight of the dead animal drove Pam to revulsion⟩ . Revulsion at slaughter cut war short [newspaper heading].

Notations and abbreviations

A	: actant
ConceptR	: Conceptual Representation
DSynt-	: deep-syntactic
ECD	: Explanatory Combinatorial Dictionary
GP	: Government Pattern
L	: a particular lexical unit
L	: given natural language
LF	: Lexical Function
LU	: lexical unit
MTT	: Meaning-Text Theory
-S	: structure
SemR	: Semantic Representation
SSynt-	: surface-syntactic

Acknowledgments

The first draft of the paper has been read (as always) by L. Iordanskaja; its present form owes much to (sometimes violent) discussions with M. Alonso Ramos, who read the subsequent text, and T. Reuther. A. Cowie and A. Grosu went through the prefinal version, and Iordanskaja checked the final one, hunting down many remaining inconsistencies. I tender [= **Oper**$_1$(*grtatitude*)] my heartfelt [= **Magn**(*gratitude*)] gratitude to all of them, while taking [= **Real**$_1$(*responsibility*)] full [= **Magn**(*responsibility*)] responsibility for all errors and obscurities that survived their scrutiny.

References

Alonso Ramos, M., and Tutin, A. (1994), 'Les fonctions lexicales du *Dictionnaire explicatif et combinatoire* pour l'étude de la cohésion lexicale', *Lingvisticae Investigationes*, 17: 161-

Alonso Ramos, M., Tutin, A., and Lapalme, G. (1992), 'Lexical Functions of *Explanatory Combinatorial Dictionary* for lexicalization in text generation', in P. Saint-Dizier and E. Viegas (edd.), *Proceedings of the 2nd Seminar on Computational Lexical Semantics* (Toulouse: IRIT), 157-68.

Becker, J. D. (1975), 'The Phrasal Lexicon', in R. Schank and B. Nash-Webber (edd.), *Proceedings of Interdisciplinary Workshop on Theoretical Issues in Natural Language Processing,* 70-73.

Cattell, R. (1984), *Composite Predicates in English* (Sydney *etc*.: Academic Press).

Dixon, R.M.W. (1991), *A New Approach to English Grammar, on Semantic Principles* (Oxford: Clarendon Press).

Gross, M. (1981), 'Les bases empiriques de la notion de prédicat sémantique', *Langage* 63: 7-52.

Iordanskaja, L., Kim, M., and Polguère, A. (1994), 'Some procedural problems in the implementation of Lexical Functions for text generation', in Wanner, L. (ed.), *Lexical Functions* (Amsterdam: Benjamins).

Ilson, R., and Mel'čuk, I. (1989), 'English BAKE Revisited (BAKE-ing an ECD)', *International Journal of Lexicography* 2: 325-45.

Mel'čuk, I. (1981), 'Meaning-Text models: A recent trend in Soviet linguistics, *Annual Review of Anthropology* 10: 27-62.

Mel'čuk, I. (1982), *Towards a Formal Language of Lnguistics* (München: W. Fink Verlag).

Mel'čuk, I. (1988a), *Dependency Syntax: Theory and Practice* (Albany, N.Y.: The SUNY Press).

Mel'čuk, I. (1988b), 'Semantic description of lexical units in an Explanatory Combinatorial Dictionary: basic principles and heuristic criteria', *International Journal of Lexicography* 1: 165-88.

Mel'čuk, I. (1989), 'Semantic primitives from the viewpoint of the Meaning-Text linguistic theory', *Quaderni di semantica* 10: 65-102.

Mel'čuk, I. (1992a), 'CHANGER et CHANGEMENT en français contemporain (étude sémantico-lexicographique)', *Bulletin de la Société de linguistique de Paris* 87: 161-223.

Mel'čuk, I. (1992b), 'Paraphrase et lexique: La théorie Sens-Texte et le *Dictionnaire explicatif et combinatoire*', in Mel'cuk *et al*. 1992: 9-58.

Mel'čuk, I. (1993), *Cours de morphologie générale* (Montréal — Paris: Les Presses de l'Université de Montréal — CNRS).

Mel'čuk, I., and Polguère, A. (1987), 'A formal lexicon in the Meaning-Text Theory (or how to do lexica with words)', *Computational Linguistics* 13: 261-75.

Mel'čuk, I., and Wanner, L. (1994), 'Lexical inheritance', in Wanner, L. (ed.), *Lexical Functions* (Amsterdam: Benjamins)

Mel'čuk, I., and Zholkovsky, A. (1984), *Explanatory Combinatorial Dictionary of Modern Russian* (Vienna: Wiener Slawistischer Almanach).

Mel'čuk, I., and Zholkovsky, A. (1988), 'The Explanatory Combinatorial Dictionary', in M. Evens (ed.), *Relational Models of the Lexicon* (Cambridge *etc*.: Cambridge University Press), 41-74.

Mel'čuk, I., *et al*. (1984, 1988, 1992), *Dictionnaire explicatif et combinatoire du français conemporain : Recherches lexico-sémantiques* I, II, III (Montréal: Presses de l'Univ. de Montréal).

Morgan, J. L. (1978), 'Two types of convention in Indirect Speech acts', in P. Cole (ed.), *Syntax and Semantics, v.9. Pragmatics* (New York *etc*.: Academic Press), 261-80.

Wanner, L., and Bateman, J. (1990), 'Lexical cooccurrence relations in text generation', in: *Proceedings of the 5th International Workshop on Natural Language Generation*, Dawson, PA.

Zolkovskij, A., and Mel'čuk, I. (1967), 'O semantičeskom sinteze [On semantic synthesis]', *Problemy kibernetiki* 19: 177-238.

Notes

[1] They are also known under a host of other names: fixed ⟨frozen⟩ phrases, idiomatic expressions, idioms, etc. I will not undertake here a terminological discussion.

[2] This is so because phrasemes cannot be studied in any one of the traditional divisions of linguistics: not in semantics nor syntax — precisely because of their non-compositional, "irregular" semantic and syntactic nature.

[3] The converse is true about English: #*Caesar Salad: As much as you like* is fully grammatical and understandable, but it is not what English speakers write on their signs .

[4] The signifier of a semantic phraseme can be constructed regularly or not; this is immaterial in the present context.

[5] The difference between the cases of the type of *black coffee* (**1b**) and those of the type of *artesian well* (**2b**) is explained by the fact that BLACK does not have in the dictionary the sense ‘without milk’ among its different senses, because it realizes this sense only with COFFEE, whereas ARTESIAN has — as its only sense — ‘[well] such that water in it rises to the surface without pumping’. In other words, the difference between cases **1b** and **2b** completely depends on the lexicographic treatment we adopt for ‘phraseologically bound’ senses. However, the problem of lexicographic description of LUs is an independent problem that has to be solved (or presupposed to be solved) prior to any discussion of phraseology.

[6] When speaking of LFs, I avoid using the term *argument* because of its multiple ambiguity.

[7] A ***main*** Surface-Syntactic Object of a lexical unit L is either its D(irect) O(bject) (if L can have a DO), or its I(ndirect)O (if L cannot have a DO), or the strongest Prep(ositional)O (in the absence of both DO and IO).

[8] A ***main*** Surface-Syntactic Object of a lexical unit L is either its D(irect) O(bject) (if L can have a DO), or its I(ndirect)O (if L cannot have a DO), or the strongest Prep(ositional)O (in the absence of both DO and IO).

[9] For these equations and for a detailed description of paraphrasing system based on LFs, see Mel’cuk 1992b; an example of paraphrasing equations is given in **3.3**.

[10] Note that *be caught in an ambush*, *fire rockets* and *spray with automatic fire* are collocations and can be described in terms of LFs: $\mathbf{Real_2}$(*ambush*) = *be caught* [*in* ART ~], $\mathbf{Real_1}$(*rocket*) = *fire* [ART ~], and $\mathbf{Labor_{12}}$(*automatic fire*) = *spray* [N *with* ~].

[11] This constraint (stipulating that using TO with the infinitive dependent on HELP implies rather indirect help than help by participation) is not strict and is often violated; many speakers use the *to*-infinitive and the bare infinitive after HELP indiscriminately.

[12] The construction "Y causes that X wants/does/ sees/..." (instead of the grammatically correct "Y causes X to want/to do/to see/...") is used in the semantic metalanguage of the ECD for semantic precision and explicitness: it allows for an explicit expression of the subject of the fact that is caused.