Igor Mel'čuk

Phraseology in the language, in the dictionary, and in the computer

Abstract: Two main families of phrasemes (= non-free phrases) are distinguished: lexical phrasemes and semantic-lexical phrasemes; the phrasemes of the first family are constrained only in their form (their meaning being free), those of the second family are constrained both in their meaning and in their form. Two basic concepts are introduced: ***compositionality*** of complex linguistic signs and the ***pivot*** of a meaning. Three major classes of phrasemes are presented: non-compositional ***idioms*** and compositional ***collocations*** and ***clichés***. A new type of general dictionary is proposed, and the lexicographic presentation of the three classes of phrasemes is illustrated. To show how the proposed approach to phraseology can be used in Automatic Language Processing, three fully-fledged examples are examined in detail.

Keywords: phraseology; compositionality; pivot (of a meaning); idioms; collocations; clichés; lexicography; automatic language processing.

Correspondence address: igor.melcuk@umontreal.ca

1 Introduction

There is no need to insist on the importance of phraseology for linguistic studies; on this point the linguistic community is in agreement. But, curiously and unfortunately, there is no agreement on either the exact content of the notion 'phraseology', nor on the way phraseological expressions should be described, nor on how they should be treated in linguistic applications, in particular, in lexicography and Natural Language Processing [= NLP]. In this article, I will try to deal with these three points: Section 2 proposes a rigorous definition of *phraseme*, a characterization of the major classes of phrasemes and an exhaustive phraseme typology, thus establishing the boundaries of phraseology; Section 3 sketches the fundamentals of the lexicographic description of phrasemes in an *Explanatory Combinatorial Dictionary*; in Section 4, three examples of difficult cases of machine translation are considered where the solutions come from the dictionary and are based on the proposed description of one of phraseme classes (namely, collocations). Finally, Section 5 summarizes the most important points of the article and formulates some paths of future research.

The theoretical framework of the discussion is Meaning-Text Theory [= MTT]. Certain of its notions and formalisms will be used without explanation. For more on MTT, please consult Mel'čuk 1981, 1988a: 43–91, 1997, 2006a: 4–11 and Kahane 2003a.

Technical terms appear, on their first mention, in **Sans-Serif Bold.**

2 Phraseology in the language

The literature on phraseology is too huge to be reviewed here even cursorily; see, for instance, the collections Everaert et al. 1995, Wanner 1996, Álvarez de la Granja 2008 and Anscombre and Mejri 2011. I will limit myself to mentioning Mel'čuk 1995 (a sketch of a theory of phraseology within the Meaning-Text framework) and the classics Bally 1909 and Weinreich 1969, which have most profoundly influenced my approach to phrasemes.

2.1 Two main families of phrasemes: lexical and semantic-lexical phrasemes

A phraseological expression, also called a *set expression*, *set phrase*, *idiomatic phrase*, *multi-word expression*, sometimes simply *idiom*, etc., is, first of all, a (**multiword**) **phrase** – that is, a linguistic expression formed by several (at least two) lexemes syntactically linked in a regular way.[1] The notorious example of an idiom *X kicks the bucket* ≈ 'person X dies of natural causes, I being flippant about X' is syntactically and morphologically structured exactly the same way as all similar phrases of the form "Transitive Verb→DirO": *kick the ball, hit John*, *squeeze her hand*, etc. (Even the expression *kick the bucket* itself can mean 'kick the bucket [full of dirty water]'.) This expression is special, i.e. phraseological, only because of its "unpredictable" meaning 'die of natural causes [said flippantly]'. A phraseological expression, or **phraseme**, is thus a phrase featuring some unpredictable properties, i.e., a **linguistically constrained** phrase, or else a phrase that is **not free**. Therefore, I have to begin with a definition of free phrase.

[1] To simplify my task, in this paper I leave aside the phrasemes of the morphological level – that is, the phraseologized combinations of morphs inside a wordform. For this family of phrasemes, or morphophrasemes, see, for instance, Beck and Mel'čuk 2011.

Definition 1: Free phrase

> A phrase is **free** if and only if [= iff] each of its lexical components L_i is selected by the speaker in a linguistically **non-constrained way** – that is, each L_i is selected strictly for its meaning and in conformity with its linguistic properties but independently of the lexical identity of other components.

In other words, while selecting L_i, the Speaker need not take into account any particular lexeme being part of the phrase in question.

Corollary: Each lexical component of a free phrase can be replaced by any of its (quasi-)synonyms without affecting its meaning and grammaticality. In the phrase *select the word freely*, you can replace any component with its synonym and the meaning is preserved: *choose the lexeme without constraint*.

Definition 2: Non-free phrase = phraseme

> A phrase is **non-free** ⟨= **phraseologized**⟩ iff at least one of its lexical components L_i is selected by the speaker in a **linguistically constrained way** – that is, as a function of the lexical identity of other component(s).

In a non-free phrase, at least one L_i is selected depending on other particular lexemes building up this phrase.

Corollary: It is not true that every lexical component of a non-free phrase can be replaced by any of its (quasi-)synonyms without affecting its meaning and grammaticality. In *kick the bucket* ≈ ‘die’ you cannot replace any of the components: *boot the bucket* or *kick the pail* do not mean ‘die’.

Let it be emphasized that the terms **(non-)constrained**, when applied to linguistic expressions, must be understood strictly in the technical sense indicated above: as selection of a lexeme regardless of the individual identity of any other lexeme of the same expression. (In the literal sense, any free phrase is of course “constrained” by the linguistic means at the disposal of the Speaker and by linguistic rules of their combination).

A phraseme violates the **freedom of selection** of its lexical components. This violation happens on the **paradigmatic axis** of speech production, as the speaker is looking (in his mental lexicon) for appropriate lexical units. The lexical selection activity proceeds in two stages:

– First, the speaker has to construct his starting meaning; for this, he selects the necessary simpler meanings and unites them into the meaning of his eventual phrase – that is, into its starting semantic representation [= SemR].
– Second, the speaker has to select the lexical units to express his starting SemR and unite them into the deep-syntactic representation [= DSyntR] of the phrase.

Accordingly, two cases of violation of selection freedom must be distinguished.

The first case. The construction of the starting meaning 'σ' [= the SemR] of the phrase L('σ') that describes the situation P is free. To obtain 'σ', the speaker selects simpler meanings 'σ_1', 'σ_2', . . ., 'σ_n' and puts them together in conformity with his own needs and general rules of his language: the language does not specifically constrain his semantic choices. But the lexical components of the phrase L('σ') cannot be chosen freely: some or all of them are selected as a function of the other. The violation of the selection freedom takes place in the transition {SemR} ⇔ {DSyntR} and manifests itself in lexical constraints. Therefore, resulting phrasemes are called **lexical**: *kick the bucket*, *pull* [N_Y's] *leg* 'lie [to N_Y] in order to have fun' or Rus. *na golubom glazu* lit. 'on blue eye' = 'pretending to act honestly and sincerely', *The rain is falling in torrents*, *It rains cats and dogs* or Rus. *Dožd′ l′ët kak iz vedra* lit. 'Rain is pouring as from bucket' and *prolivnoj dožd′* 'torrential rain' are typical **lexical phrasemes**.

Definition 3: lexical phraseme

> A phraseme L('σ') is **lexical** iff its meaning 'σ' is constructed by the Speaker freely, but its lexical components L_i (all or some) are selected in a **constrained way**.

The second case. Not only the lexical composition of the phraseme is constrained, but also its meaning. To describe the situation P, the Speaker is forced by the language to select the starting meaning 'σ', and he can take no equivalent meaning. Thus, the phrase L('σ') is constrained semantically **and** lexically. This type of phraseme is thus "doubly" constrained: in the transition {ConceptR} ⇔ {SemR} (semantic constraints) and then in the transition {SemR} ⇔ {DSyntR} (lexical constraints). This is a **semantic-lexical phraseme**. A simple example is the sign *Wet paint*: Russian says in this context *Ostorožno, okrašeno* lit. 'Caution, painted' rather that #*Syraja kraska* 'Wet paint' or even #*Ostorožno, **vy**krašeno* (with a different aspect prefix); and in English it would be inappropriate to write on a sign #*Caution, painted*, although this is a perfectly grammatical and semantically correct phrase (the symbol "#" indicates pragmatic unaceptability). Here the language prescribes the meaning to express and its specific lexical expression.

Definition 4: semantic-lexical phraseme

> A phraseme L('σ') is **semantic-lexical,** iff not only the components L_i of its lexical expression, but also the components 'σ_i' of its meaning are selected by the Speaker in a **constrained way**.

Examples: *in other terms/in other words*; *to make a long story short*; Rus. *inače govorja* lit. 'speaking differently', *koroče govorja* lit. 'speaking shorter' or *čto i trebovalos′ dokazat′* 'Q.E.D.'.

Thus, a major partition splits phrasemes into two subsets: **lexical phrasemes** and **semantic-lexical phrasemes.**

2.2 Compositionality and semantic pivot

To develop a finer typology of phrasemes, two auxiliary notions are needed: compositionality of linguistic signs and the semantic pivot of a meaning.

2.2.1 Compositionality of complex linguistic signs

A **linguistic sign** s is a triplet

$$\mathbf{s} = \langle\ \text{‘}\sigma\text{’}\ ;\ /s/\ ;\ \Sigma\ \rangle, \text{ where:}$$

‘σ’ is the **signified**, or informational content, most often a linguistic meaning;
/s/ is the **signifier**, or a physical signal, most often a string of phonemes or of characters;
Σ is the **syntactics**, or a set of data specifying the cooccurrence of **s** with other signs. (See Mel’čuk 1982: 40-51 and 2006b: 384-386.)

For instance, the noun AIRCRAFT is represented as a linguistic sign like this:

⟨ ‘vehicle designed to fly’-SG/PL ;[2] /ɛ́ᵊrkræft/ ; Σ = noun, countable, Lexical Functions: *land*$_{(V)}$, *crew*, . . . ⟩

Simple signs and elements of signs combine into complex signs by the operation of **linguistic union** ⊕. For a particular language, this operation is represented by a set of linguistic rules that tell us how, in this language, signs must be united:

- the signifieds are united by putting the SemR of an argument into the corresponding argumental position of “its” predicate;
- the signifiers are united by juxtaposing the strings of phonemes and applying all necessary morphological operations to the strings of phonemes;
- the syntactics are united by retaining the combinatorial data valid for the resulting complex sign.

To simplify the presentation, in what follows, I will define two-element complex signs and two-element phrasemes; the proposed definitions can be easily generalized to include any number of components within a complex sign or a phraseme.

Definition 5: compositional complex linguistic sign
A complex linguistic sign **AB** is **compositional** iff **AB** = **A** ⊕ **B**.

This means that, for the sign **AB** = ⟨ ‘AB’ ; /AB/ ; $\Sigma_{\mathbf{AB}}$ ⟩, its signified ‘AB’ = ‘A’ ⊕ ‘B’, its signifier /AB/ = /A/ ⊕ /B/ and its syntactics $\Sigma_{\mathbf{AB}} = \Sigma_{\mathbf{A}} \oplus \Sigma_{\mathbf{B}}$.

[2] This notation indicates that the meanings of the grammemes SINGULAR and PLURAL are part of the signified of the radical of the lexeme AIRCRAFT.

From Definition 5, it follows that compositionality is an absolute notion, which does not admit degrees: a complex sign is compositional or not. Compositionality concerns the three components of the sign independently; in what follows I will consider only the compositionality of signifieds, i.e., **semantic compositionality**.

A free phrase is necessarily compositional: it is thanks to this property of free phrases that linguistic communication is possible. To master language **L** means to have in the brain simple signs of **L** and the rules of the operation ⊕ for **L**.

The selection of lexical units happens on the **paradigmatic axis** of language while their combination implies the **syntagmatic axis**. Taking into account the two axes of speech production guarantees that our characterization of phrasemes is exhaustive.

2.2.2 The semantic pivot of a meaning

The notion of semantic pivot is needed for a finer characterization of non-compositional phrasemes.

Definition 6: semantic pivot (of a meaning)

Let there be meaning ‘σ’ that is divided into two parts, ‘σ_1’ and ‘σ_2’ (‘σ’ = ‘σ_1’ ⊕ ‘σ_2’).
The part ‘σ_1’ of meaning ‘σ’ is called the **semantic pivot** of ‘σ’ iff the other part ‘σ_2’ is a predicate of which ‘σ_1’ is the argument: ‘σ’ = ‘σ_2’(‘σ_1’).

The semantic pivot of meaning ‘σ’ is logically different from the **communicatively dominant component** of ‘σ’, which is the minimal paraphrase of ‘σ’ (Mel’čuk 2001: 29–43). Thus, in the meaning of the phraseme *take a shower* ‘wash oneself under a shower’, the semantic pivot is ‘shower’, while the communicatively dominant component is ‘wash’. The semantic pivot will be identified in the examples by shading. Note that the semantic pivot of a multi-word expression **E** does not have to coincide with the lexical meaning of one of **E**’s components. Thus, in the phrase *sea dog* ‘person very experienced in navigation at sea’ the semantic pivot ‘person’ is carried by none of its two components.

The notion of semantic pivot will be used to sharpen the typology of idioms, see below, 2.3.1.

2.3 Major classes of phrasemes

Crossing the two dimensions – being subject to lexical *vs*. semantic-lexical constraints and being compositional *vs*. non-compositional – gives four major classes of phrasemes. However, one of these logically possible classes, namely semantic-lexical non-compositional phrasemes, cannot exist: if a non-free phrase is non-compositional, it has, by definition, a “holistic” meaning that is associated with it as a whole; therefore, this meaning cannot be constructed by the speaker for the occasion; therefore, it does not make sense to talk about constrained/non-constrained character of its construction. As a result, a natural language has just three major classes of phrasemes: **idioms**, **collocations** and **clichés**, as in Figure 1.

Compositionality of phrasemes Nature of constraints	non-compositional	compositional
lexical	IDIOMS	COLLOCATIONS
semantic-lexical	impossible	CLICHÉS

Figure 1. Three major classes of phrasemes

2.3.1 Idioms

Idioms constitute the most known and best studied subset of phrasemes. Three subclasses of idioms can be distinguished: full idioms, semi-idioms and quasi-idioms.

Definition 7: idiom

A lexical phraseme is an **idiom** iff it is non-compositional.

An idiom is indicated in print by elevated half-brackets: ⌜ . . . ⌝.

Examples: ⌜*cheek by jowl*⌝ ‘in close association’, ⌜*The game is up*⌝ ‘The deceit is exposed’, ⌜*come to* [N_X’s] *senses*⌝ ‘X becomes conscious again’, ⌜*put* [N_Y] *on the map*⌝ ‘make the place Y well-known’, ⌜*bull session*⌝ ‘long informal talk on a subject by a group of people’, ⌜*game of chicken*⌝ ‘showdown between two opponents where none is disposed to yield and both lose if they push the conflict to the end’, Rus. ⌜*ostat′sja s nosom*⌝ lit. ‘X remains with nose’ ≈ ‘X gets nothing in a situation where X was supposed to obtain something’, ⌜*sinij čulok*⌝ ‘bluestocking’, etc.

An idiom can be characterized by the degree of its **transparence/opacity**: the degree to which its meaning includes the meanings of its components. Three types of idioms can be distinguished in such a way: **full idioms**, **semi-idioms** and **quasi-idioms**. All of them are non-compositional, but the degree of their transparency varies.

Definition 8: full idiom

An idiom **AB** is a **full idiom** iff its meaning does not include the meaning of any of its lexical components: ‘AB’ ⊅ ‘A’ **and** ‘AB’ ⊅ ‘B’.

Examples: ⌜*put*$_{\mathrm{A}}$ [N$_{\mathrm{Y}}$] *through its paces*$_{\mathrm{B}}$⌝ ‘to test Y thoroughly’, ⌜*go*$_{\mathrm{A}}$ *ballistic*$_{\mathrm{B}}$⌝ ‘suddenly become very angry’, ⌜*by*$_{\mathrm{A}}$ *heart*$_{\mathrm{B}}$⌝ ‘remembering verbatim’, ⌜*bone*$_{\mathrm{A}}$ *of contention*$_{\mathrm{B}}$⌝ ‘reason for quarrels or fights’, Rus. ⌜*jabloko*$_{\mathrm{A}}$ *razdora*$_{\mathrm{B}}$⌝ lit. ‘apple of discord’ = ‘bone of contention’, ⌜*delat′*$_{\mathrm{A}}$ *nogi*$_{\mathrm{B}}$⌝ lit. ‘do legs’ = ‘flee’, ⌜*polezt′*$_{\mathrm{A}}$ *v butylku*$_{\mathrm{B}}$⌝ lit. ‘try.to.get into bottle’ = ‘stubbornly insist on something in a dangerous situation’, etc.

Definition 9: semi-idiom

An idiom **AB** is a **semi-idiom** iff its meaning 1) includes the meaning of one of its lexical components, but not as its semantic pivot, 2) does not include the meaning of any other component and 3) includes an additional meaning ‘C’ as its semantic pivot:

‘AB’ ⊃ ‘A’, **and** ‘AB’ ⊅ ‘B’, **and** ‘AB’ ⊃ ‘C’.

Examples: ⌜*private*$_{\mathrm{A}}$ *eye*$_{\mathrm{B}}$⌝ ‘private detective’, ⌜*sea*$_{\mathrm{A}}$ *anemone*$_{\mathrm{B}}$⌝ ‘predatory polyp living in the sea’, Rus. ⌜*mozolit′*$_{\mathrm{B}}$ *glaza*$_{\mathrm{A}}$⌝ ‘be too often or for too long before Y’s eyes’ (lit. ‘make corns on Y’s eyes’).

Thus, a semi-idiom is semi-transparent (or semi-opaque, depending on whether you are an optimist or a pessimist).

Definition 10: quasi-idiom (= weak idiom)

An idiom **AB** is a **quasi-idiom**, or **weak idiom**, iff its meaning 1) includes the meaning of both of its lexical components, neither as the semantic pivot, and 2) includes an additional meaning ‘C’ as its semantic pivot: ‘AB’ ⊃ ‘A’, **and** ‘AB’ ⊃ ‘B’, **and** ‘AB’ ⊃ ‘C’.

Examples:

⌜*start*$_{\mathrm{A}}$ *a family*$_{\mathrm{B}}$⌝	‘conceive the first child with one’s spouse, thereby starting a full-fledged family’
⌜*barbed*$_{\mathrm{A}}$ *wire*$_{\mathrm{B}}$⌝	‘artifact designed to make obstacles with – wire with barbs fixed on it at small regular intervals’
Fr. ⌜*donner*$_{\mathrm{A}}$ *le sein*$_{\mathrm{B}}$ [*à* N$_{\mathrm{Y}}$]⌝ ‘breast-feed Y’	‘feed the baby Y by putting the teat of a breast into the mouth of Y’

2.3.2 Collocations

Collocations have attracted the attention of linguists much later than idioms and are still not sufficiently theorized nor inventorized.

Definition 11: collocation

A lexical phraseme is a **collocation** iff it is compositional.

Examples: *heavy* ACCENT, Rus. *sil′nyj* AKCENT lit. ‘strong accent’, Fr. ACCENT *à couper au couteau* lit. ‘accent to cut with the knife’; *soundly* ASLEEP, Rus. SPAT′

glubokim snom lit. ‘asleep with deep sleep’; *ARMED to the teeth*; *fasten* ⟨= *buckle up*⟩ *the SEATBELT*, Rus. *zastegnut′* ⌜*PRIVJAZNOJ REMEN*⌝ lit. ‘button.up seatbelt’; *leap YEAR*, Rus. *visokosnyj GOD* (the adjective VISOKOSNYJ is used only with GOD ‘year’).

Interestingly, a collocation is binary – it consists of two major elements: a **base**, lexical expression chosen freely by the speaker (shown in *SMALL CAPS*), and a **collocate**, lexical expression chosen as a function of the base to express a given meaning bearing on the base. (The base and the collocate can each consist of several lexemes.) A collocation is semantically compositional, since its meaning is divisible into two parts such that the first one corresponds to the base and the second to the collocate. The meaning of the base is always the semantic pivot of the collocation. For more on collocations in the Meaning-Text framework, see Mel’čuk 2003a, 2003b and 2004, as well as Alonso Ramos 2004a, b, c and Vincze et al. 2011.

This should not be understood as implying that a collocate – taken as such, outside the collocation – necessarily has the meaning it expresses within the collocation. Thus, in the collocation *sit for an exam* ‘undergo an exam’, the verb SIT expresses the meaning ‘undergo’; but in an English dictionary, the verb SIT does not have to carry this meaning: ‘undergo’ is not its inherent, but context-imposed signified.

In English, you ***make*** *a decision*, and in Britain, you can also *take* it. For the same thing, French says *prendre* [= ‘take’] *une décision*, German – *eine Entscheidung* ***treffen/fällen*** [= ‘meet/fell’], Russian – ***prinjat′*** [= ‘accept’] *rešenie*, Turkish – *karar* ***vermek*** [= ‘give’], Polish – ***podjąć*** [= ‘take up’] *decyzję*, Serbian – ***doneti*** [= ‘bring’] *odluku*, and Korean – *gyeoljeongeul* ***haerida/naerida*** [= ‘do/put.down’]. This clearly shows that boldfaced verbs are selected as a function of the noun meaning ‘decision’. If instead of DÉCISION a French speaker uses CHOIX ‘choice’ (*Jean a pris la décision de rester* ‘Jean has taken the decision to stay’ ≅ *Jean a . . . le choix de rester* ‘Jean has . . . the choice to stay’), he has to say FAIRE ‘make’ rather than PRENDRE ‘take’: *Jean a fait* ⟨**a pris*⟩ *le choix de rester* ‘Jean has made the choice to stay’.

Collocations are extremely variegated and very numerous in any language (in the millions). Two major types are distinguished: standard and non-standard collocations.

Definition 12: standard collocation

A collocation “Base–**r**–Collocate” is **standard** iff the following two conditions are satisfied:

1) Semantic relation **r** is applicable, in language **L**, to many different bases and defines many different collocates.

2) Collocations with the semantic relation **r** between their components participate in paraphrasing.

In other words, **L** has many collocations where the relation between the base and the collocate is **r**; and "many" means here at least several dozen or, better, hundreds. In the examples below, the base of collocation is in small caps.

Examples: 'John apologizes to Mary': *John makes ⟨ = offers⟩ Mary an APOLOGY. ~ Mary receives an APOLOGY from John.* | 'John despairs': *John is in DESPAIR. ~ John is DESPERATE. ~ DESPAIR seized* ⟨≈ *overcame*⟩ *John.* | 'John supports Mary': *John lends ⟨ = gives⟩ Mary his SUPPORT. ~ John throws his SUPPORT behind Mary. ~ John is very SUPPORTIVE of Mary. ~ Mary has John's SUPPORT.* ~ [*Significant*] *SUPPORT comes to Mary from John.*

Definition 13: non-standard collocation

A collocation "Base–**r**–Collocate" is non-standard iff the following two conditions are satisfied:

1) Semantic relation **r** is applicable, in language **L**, only to a few different bases (in the minimal case, to one base) and defines a few different collocates (again, minimally, just one).

2) Collocations with the semantic relation **r** between their components do not participate in paraphrasing.

Examples: *leap YEAR*, **r** = 'having 366 days'; *black COFFEE*, **r** = 'with no dairy product added'; *LAUGH in* [N_X's] *sleeve*, **r** = 'trying to hide the fact of . . .'; *spiked HEELS*, **r** = 'long and thin'; etc.

2.3.3 Clichés

Clichés are the latest arrival on the phraseological scene. The most numerous among phrasemes, they are also the most difficult to pinpoint and collect and, at the current time, the least studied.

Definition 14: cliché

A semantic-lexical phraseme is a **cliché**.

Examples: *If you've seen one, you've seen them all*; *Happy birthday to you!*; *no matter what*; *We all make mistakes*; *Will you marry me?*; etc.

A cliché is a compositional expression – just as a collocation is – used for a complex meaning 'σ' that language **L** prescribes to use for the description of a given situation P – to the exclusion of all other equivalent meanings. Thus, in English we ask *What is your name?* and answer *My name is* [N] or *I am* [N]; Russians say *Kak vas zovut?* lit. 'How do they call you?' and *Menja zovut* [N] 'They call me [N]'. The sentences *Kak vaše imja?* and *Ja* [N], the literal renderings of the English expressions, are fully understandable and grammatical, but not standard.

A cliché is characterized by a **lexical anchor** (or anchors), which is the lexeme whose meaning identifies the use of the cliché: *What is your name?* and *Kak vas zovut?* 'What do the call you?' have NAME/IMJA as their anchor. (As we see in *Kak vas zovut?*, a cliché's lexical anchor does not have to be explicitly present in the cliché.) In a dictionary, clichés are described under their lexical anchors.

Clichés are further subdivided in two major classes: pragmatically constrained clichés (= pragmatemes) and pragmatically non-constrained, or "normal," clichés.

– **Pragmatemes**. Along with lexical and semantic-lexical constraints that violate the freedom of lexical selection on the paradigmatic axis, natural language features a third type of constraint – situational, or **pragmatic**, constraints. Such constraints stipulate that a particular cliché may be required by a particular situation of its use. (Note that what is meant is the **situation of the use** of a cliché, not the **situation described** by the cliché.) Thus, as a warning on a container of perishable food, English says *Best before* . . ., while in Russian, this will be *Srok godnosti* . . . lit. 'Term of.validity . . .', in Polish, *Najlepiej spożyć* . . . lit. 'The best [is] to consume . . .', in French, *À consommer avant* . . . lit. 'To consume before . . .', and in German, *Mindestens haltbar bis* . . . lit. 'At.least keepable till . . . '. All these expressions are fully constrained and compositional – that is, they are clichés. But this is a particular type of clichés, since they are used in a very particular situation: as an official statement [**on a container of perishable food manufactured for sale**]. The boldfaced indication in brackets is a pragmatic constraint on this particular cliché.

Pragmatic constraints are in principle applicable to any type of lexical expression – not only to phrasemes but to lexemes as well; here are examples:

pragmatically constrained

idioms	: *Break a leg!* [**to a performer who is going on stage**]
collocations	: *Wet paint* [**on a sign**]
clichés	: *No parking* [**on a sign**]
lexemes	: *Roger!* 'I understood' [**in a radio communication**] Pol. *Smacznego!* lit. 'Of tasty!' = 'May your food be tasty' [**to people starting a meal**]

However, among pragmatically constrained lexical expressions, clichés occupy a special place: a crushing majority of pragmatically constrained phrasemes are clichés. Therefore, it is convenient to give pragmatically constrained clichés a special name: **pragmatemes**.

Definition 15: pragmateme

A pragmatically constrained cliché is a **pragmateme**.

Examples: *Hold the line!* [**in a telephone conversation**], *Watch your step!* [**on a sign**], *X – all you can eat* [**on a sign in a restaurant**], *Emphasis mine* [**after a quotation in a written text**], *Return to sender* [**on a postal sending**], *Who's there?* [**answering a knock on the door**], etc. (Such a cliché as *What's your name?* is not a pragmateme: it can be used in any situation; likewise, *Sorry to keep you waiting*, *I am in the mood* [*for* Y], *Would you mind* [Y-*ing*?], *It's a proven fact*, etc.)

– **Pragmatically non-constrained clichés**. There is not much to say about them, except that they include two special subgroups:

1) Compositional proverbs and sayings, such as *A watched pot never boils*; *Money isn't everything*; *Worrying never did anyone any good*; . . . The linguistic meaning of a proverb is its literal meaning, which corresponds, in an idiosyncratic way, to its informational, or conceptual, content. The proverb *A watched pot never boils* semantically means what it says, but to use it properly you have to know that you say it to express the following conceptual content: «If you are waiting for something and are nervous, it will take longer».
2) Compositional **complex proper names**, such as *The Old/New Testament*, *A Midsummer Night's Dream* [Shakespeare's play], *The Moonlight Sonata* [Beethoven's piano sonata], *City of Lights* [nickname of Paris], *Eternal City* [nickname of Rome], *Red Planet* [nickname of Mars], . . . (Bosredon 2011). Again, the linguistic meaning of a complex proper name is its literal meaning, but it idiosyncratically corresponds to one particular referent.

2.4 Typology of phrasemes

I can now present all the major subclasses of phrasemes and their taxonomy as in Figure 2.

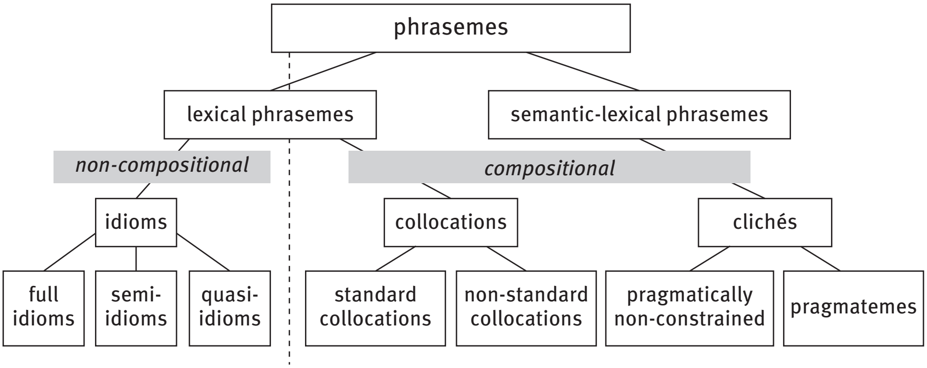


Figure 2. Phraseme typology

3 Phraseology in the dictionary

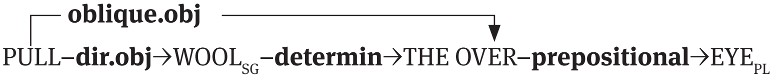
The dictionary considered here is the *Explanatory Combinatorial Dictionary* [= ECD]; its principles, structure and basic notions are taken for granted (see Mel'čuk and Zholkovsky 1984, Mel'čuk 1988b, Mel'čuk et al. 1984–1999, Mel'čuk et al. 1995, Mel'čuk 2006a, Mel'čuk and Polguère 2007). I will discuss only the lexicographic presentation of phrasemes.

3.1 Lexicographic presentation of non-compositional phrasemes (= idioms)

An idiom is a **lexical unit**, just as a lexeme is one. Idioms are, then, described in an ECD the same way as lexemes: each has its proper dictionary article, featuring the same structure as a lexeme article, with one important difference: since an idiom is a multiword phrase, it is supplied with its SSynt-tree. For instance:

⌜PULL THE WOOL OVER [N_Y's] EYES⌝, verbal idiom

Definition
X pulls the wool over Y's eyes: 'X tries to deceive Y in order to hide from Y what X is really doing'.
The surface-syntactic structure


oblique.obj
PULL–dir.obj→WOOL$_{SG}$–determin→THE OVER–prepositional→EYE$_{PL}$

Government Pattern

X ⇔ I	Y ⇔ II	
1. N	1. *of* N	THE←**determ**–EYES–**attrib**→OF N
	2. N's	EYES–**possessive**→N's
	3. $A_{(poss)}$(N)	EYES–**determ**→$A_{(poss)}$(N)

Don't pull the wool over foreigners' eyes! | *He tried to pull the wool over my ⟨John's⟩ eyes*.

The number of idioms in a particular language is probably around 10,000; thus, the English idiom dictionary of Cowie et al. 1993 contains about 7,000 idiomatic expressions, and the French idiom dictionary Rey and Chantreau 1993, about 9,000; an excellent Russian-English dictionary of idioms (Lubensky 1995) presents some 13,000 idiomatic units.

3.2 Lexicographic presentation of compositional phrasemes (= collocations and clichés)

Compositional phrasemes – collocations and clichés – are not lexical units; they do not have their own dictionary articles and are described in the articles of their bases/anchors. For instance, the collocation *ARMED to the teeth* does not have a separate entry, but appears under ARMED; Rus. *vypisat′ ČEK* lit. ‘write out check’ = ‘draw a check’ is given under ČEK; and so forth. The same is true of clichés.

3.2.1 Lexicographic presentation of collocations

The number of collocations in languages of *Standard Average European* type is very high: no less than ten times the number of lexemes, which means millions. Therefore, the lexicographic description of collocations requires a special formal apparatus that would allow for their elegant systematic presentation in the dictionary and, at the same time, for facilitating automatic processing. Such an apparatus is the system of **lexical functions** [= LFs]. It is impossible to introduce here the notion of LF or offer the reader a detailed review thereof (see Žolkovskij and Mel’čuk 1967, Mel’čuk 1974: 78–109, 1982a, 1996, 2003a, b, Wanner (ed.) 1996, Kahane 2003b, Kahane and Polguère 2001, Alonso Ramos 2005, 2010). I will limit myself to a few examples of **standard** and **non-standard** LFs, used for the description of, respectively, standard and non-standard collocations – in order to show afterwards how LFs can be exploited in NLP.

Standard collocations described by standard LFs

A standard LF $\mathbf{f}_{\text{stand}}$ describes a family of standard collocations where the semantic relation between the base and the collocate is institutionalized in the language; an $\mathbf{f}_{\text{stand}}$ specifies this relation simply by its name. In conformity with the definition of standard collocations, an $\mathbf{f}_{\text{stand}}$ is actively involved in deep-syntactic paraphrasing, see Section 4.

- **Verbal Standard LFs**
  – **Support verbs**

	RESPONSIBILITY	**CARE**	**ACCUSATION**	**AID**
	X’s ~ concerning Y	**X’s ~ concerning Y**	**X’s ~ of Y of Z**	**X’s ~ to Y**
Oper$_1$	*carry* [ART ~]	*give* [~ to N_Y]	*level* [ART ~ at N_Y]	*come* [to the ~(→of N_Y)]
Func$_2$	~ *includes* [N_Y]	~ *is aimed* [at N_Y]	~ *weighs* [on N_Y]	~ *comes* [to N_Y from N_X]
Labor$_{12}$	——	*surround* [N_Y with ~]	*bring* [N_Y under ~]	*support* [N_Y with N_X’s ~]

– **Realization verbs**

	PRIZE	DOCTOR	TRAP	ASPHALT
	X's ~ to Y for Z	~ X of Y	X's ~ for Y	~ used by X on Y
Real$_2$	*win* [ART ~]	*see* [ART ~]	*fall* [into ART ~]	——
Fact$_2$	~ *goes* [to N$_Y$]	~ *sees* [N$_Y$]	~ *catches* [N$_Y$]	~ *covers* [N$_Y$]
Labreal$_{12}$	*honor* [N$_Y$ with ART ~]	——	*catch* [N$_Y$ with ART ~]	*cover* [N$_Y$ with ~]

- **Adjectival Standard LFs (intensifiers/mitigators)**

	WET	DRUNK	BREATHE	ROLE	LAUGHTER
Magn	*~ to the bone*	*dead, stone ~, ~ as a skunk; //smashed*	*~ heavily*	*important < crucial < critical ~*	*hysterical, side-splitting ~ ; uncontrollable ~*

	WOUND	DRUNK	BREATHE	ROLE	LAUGHTER
Anti-Magn	*light ~; //scratch*	*slightly ~; //tipsy*	*~ lightly*	*small, secondary ~*	*muffled ~*

Non-standard collocations described by non-standard LFs

A non-standard LF $\mathbf{f}_{\textbf{non-stand}}$ describes a non-standard collocation where the semantic relation between the base and the collocate is not institutionalized in the language; to specify this relation, $\mathbf{f}_{\textbf{non-stand}}$ must be described in the same metalanguage as that used for lexicographic definitions:

with no diary products added(COFFEE)	: *black* [~]
drinking up the glass at one go(DRINK)	: [~] *bottoms up*
used too much(EXAMPLE)	: *hackneyed* [~]

Non-standard LF do not participate in paraphrasing.

To illustrate the lexicographic description of collocations, here is a lexical entry for the noun BATTLE (as in *Fierce battles are raging within 25 miles of Tangkin*):

BATTLE, NOUN, COUNTABLE

Definition

Battle between X and Y for Z: 'Armed confrontation between group X and group Y for Z'.

Government Pattern

X ⇔ I	Y ⇔ II	Z ⇔ III
1. *of* N	1. *with* N	1. *for* N
2. N's	2. *against* N	2. *over* N
	3. *between* N *and* N	3. *to* V_{inf}
	4. between N_{PL}	

a battle of Philippino guerrillas/their battle with ⟨*= against*⟩ *the Japanese; battles between Palestinian factions for* ⟨*= over*⟩ *the border control* ⟨*= to control the border*⟩

Lexical Functions

$\text{Syn}_{\cap}$	: *engagement* < *combat; action; fight; firefight*
$\text{V}_{0\cap}$	: *battle*$_{V}$
$\text{S}_{1/2}$	: *combatant; adversary, enemy*
S_{loc}	: *battlefield, battleground*
Mult	: //*hostilities; war*
Loc_{in}	: *in* [~]
Ver	: *winning* [*a winning* ~]
AntiVer	: *losing* [*a losing* ~]
Magn	: *pitched; ferocious, fierce, grueling, intense, rude, violent; bloody* < *murderous* [huge losses] < *mortal; royal* | postposed
AntiMagn	: //*skirmish*
Oper_1	: *fight* [ART ~]; *be, be locked, be engaged* [*in* ART ~ *against* N_Y]
IncepOper_1	: *engage* [ART ~]
ContOper_1	: *continue* [ART ~]
FinOper_1	: *stop* [ART ~]
CausOper_1	: *send* [N_X *in* ~]
[Magn+Func_0]	: *rages*
Func_{1+2}	: *opposes* [N_X *to* N_Y; N_X *and* N_Y], *pits* [N_X *against* N_Y]
nonFunc_0	: //*guns are silent*
IncepFunc_0	: *breaks out*
IncepLabor12	: //*engage* [N_Y]
Real_1	: *win* [ART ~]
AntiReal_1	: *lose* [ART ~]
Son	: *rumbles*
X and Y being individuals in physical contact	: *close-quarter*, *hand-to-hand* [~]
X and Y being ships	: *naval* [~]
X and Y being planes	: *aerial, air* [~] //*dogfight*
X and Y being of quite unequal forces	: *unequal*; *see-saw* [~]
X and Y being of rather equal forces	: *tight* [~]
more difficult for X	: *up-hill* [~]
X's first B.	: //*baptism of fire*
X begins to participate in B.	: *joins* [*the* ~]

The lexicographic description of collocations presents an additional problem related to a high level of redundancy: on many occasions, very numerous bases have the same collocate – as, for instance, all names of feelings have the same value of the lexical function Oper$_1$: *feel anger*, *gratitude*, *joy*, *pity*, *shame*, *sorrow*, etc. (of course this happens because of feeling names' semantic relatedness, which ensures partial motivation for the selection of the collocate). A solution is the generalization of collocation descriptions along several possible axes (see, for instance, Mel'čuk and Wanner 1996).

3.2.2 Lexicographic presentation of clichés

Being compositional, the meaning of a cliché need not be indicated in the dictionary; what has to be indicated is the conceptual (= informational) content to which a given cliché corresponds (a conceptual representation is printed in Monaco and underlined). Thus, for the content « I want you to tell me your name », English says *What's your name?*, while in Russian the corresponding expression is *Kak vas zovut?* lit. 'How do they call you?' and in German, *Wie heissen Sie?* lit. 'How are called you?', which have different linguistic meanings.

Clichés (including pragmatemes) and pragmatically constrained lexemes are presented in the articles of their anchor(s) in a way similar to non-standard LFs, except that instead of the description of their linguistic meaning, the dictionary gives a description of their conceptual content. For instance:

PAY$_{(V)}$	
« without having to pay »	: *free of charge*
LATE	
« Even if this is happening later than needed, it is OK »	: *Better late than never*
PUBLISH	
« [the text in question] is supposed to be published shortly »	: *To appear* [**in a bibliographic reference**]
DOG	
« There is an aggressive dog on premises »	: *Beware of (the) dog* [**on a sign**]

3.3 New type of general dictionary

The proposed lexicographic description of phrasemes entails a new concept of general dictionary. Traditionally, a dictionary is a huge list of words supplied with all types of necessary or useful information. But if the dictionary also has to store and systematically describe all the phrasemes, which outnumber words at least 10 to 1, it ceases to be a dictionary of words; it becomes a dictionary of phrases or, more precisely, of minimal phrases – that is, phrases that cannot be fully represented in the lexicon in terms of other, smaller phrases. The idea that what is actually needed is a dictionary of multiword expressions was put forth in a concise article Becker 1975; coming from a different direction (language teaching), Nattinger 1980 also underscored the necessity of a “phrasal” dictionary. Bogusławski and Wawrzyńczyk 1993 and Bogusławski and Danielewiczowa 2005 constitute an excellent illustration of what such a dictionary should look like: their dictionary includes idioms, collocations and clichés, but also syntactic constructions (for instance, «N_X *of* $N_{(period)Y}$» ‘X who/which became well known during period Y’: *book of the year* or *cover girl of the month*). More recently, many dictionaries of idioms and collocations have been published for different languages, but what I am aiming at here is a general dictionary where words and multiword expressions are stored and described together and in parallel. The ECD is intended to be such a dictionary; a first attempt at presenting a reduced model of an ECD for Spanish is Alonso Ramos 2004a and for French, Mel’čuk and Polguère 2007.

4 Phraseology in Natural Language Processing

Idioms and clichés must be listed in the dictionary, and I have shown how this could be done. But the collocations pose a serious problem for automatic processing, in particular for machine translation, given their number and variety. Lexical functions offer a reasonable solution.

LFs can be used in NLP – in particular, in machine translation and text generation – in two ways. On the one hand, all LFs ensure correct lexical selection when translating the collocations of the type (English-Russian) ***grave*** *illness* ~ ***tjažëlaja*** *bolezn′* lit. ‘heavy illness’, ***put*** [N_Y] *in danger* ~ ***podvergat′*** [N_Y] *opasnosti*$_{DAT}$ lit. ‘submit [N_Y] to danger’ or ***take*** *flight* ~ ***obratit′sja*** *v begstvo* lit. ‘turn.oneself in flight’. All such “exotic” equivalences are covered by pairs of ECD-type dictionaries; LFs, being linguistically universal, play the role of an interlingua.

On the other hand, standard LFs underlie paraphrasing at the deep-syntactic level. This paraphrasing is necessary, among other things, to resolve syntactic mismatches between the input and output sentences S_{source} and S_{target}, such mismatches being extremely frequent in parallel texts. Only paraphrasing can allow a translation system to construct an acceptable deep-syntactic structure for the output sentence S_{target} in the case of a serious mismatch between the vocabulary of S_{target} and its DSyntS, “inherited” from S_{source}. Thus, consider the sentence (1a) and its translations in Russian and French (1b-c):

(1) a. *She competes internationally.*
b. Rus. *Ona učastvuet v meždunarodnyx sorevnovanijax*
‘She participates in international competitions’.
c. Fr. *Elle participe à des compétitions internationales* [idem].

The verb meaning ‘[*to*] compete’ (in the needed sense) does not exist in Russian or French. However, a verb V can – under appropriate conditions – be paraphrased by the deverbal noun $\mathtt{S}_0(\mathrm{V})$ and one of its support verbs: $\mathrm{V} \Leftrightarrow \mathtt{S}_0(\mathrm{V})\leftarrow\mathbf{II}-\mathtt{Oper}_1(\mathtt{S}_0(\mathrm{V}))$. This formula describes all equivalences of the type *compete* ⇔ *participate in competition(s)*; the noun *competition* has direct equivalents in Russian and French.

For a universal DSynt-paraphrasing system, see Žolkovskij and Mel’čuk 1967, Mel′čuk 1974: 141–176, 1988c, 1992, 2004, and Milićević 2007: 245–333; Mel’čuk and Wanner 2006 deals specifically with the problem of syntactic mismatches in machine translation; the use of LFs in text generation is described in Iordanskaja et al. 1996 and Lareau and Wanner 2007. A paraphrasing system for Russian has been implemented and tested in a series of computer experiments: Apresjan and Cinman 1998 and 2002.

I will now present three examples of translation that are difficult because of the collocations involved, in order to show how the use of LFs ensures good results.

Example 1: The verb STRIKE

Take the sentence in (2a) and its closest (= most literal) Russian translation in (2b):

(2) a. *The book thief struck again.*
b. *Knižnyj vor snova soveršil kražu* lit. ‘Book thief again committed theft’.

It is impossible to translate STRIKE in (2a) as UDARJAT′ ‘strike’: the result would be incomprehensible. The correct choice is the collocation *soveršit′ kražu* ‘commit a theft’. But where and how can we establish the equivalence *strike* ≡ *soveršit′ kražu*? In different contexts, the verb STRIKE has lots of other equivalents in Russian:

(3) a. *The hurricane **struck** the island again.* ≡
*Uragan snova **obrušilsja** na ostrov* lit. ‘Hurricane again fell.down on island’.
b. *The bullet **struck** him in the shoulder.* ≡
*Pulja **popala** emu v plečo* lit. ‘Bullet hit him in shoulder’.
c. *A suicide bomber **struck** in the market.* ≡
*Terrorist-smertnik **podorval sebja** na rynke* lit. ‘Suicide bomber exploded himself in market’.

And so forth.

However, if we think of LFs, the answer comes immediately: all the illustrated uses of STRIKE are values of LF Fact$_0$ since Fact$_0$(L) ≈ 'perform the action that (the denotation of) L is supposed to perform in conformity with its nature'. A Russian ECD must have:

Fact$_0$(VOR 'thief')	: *krast′* 'steal', *soveršat′ kražu* 'commit a theft'
Fact$_0$(URAGAN 'hurricane')	: *obrušit′sja* [*na* N] 'strike [N]'
Fact$_0$(PULJA 'bullet')	: *popast′* 'hit'
Fact$_0$(TERRORIST-SMERTNIK 'suicide bomber')	: *podorvat′ sebja* 'explode himself'

An English ECD gives the same indications for the above uses of STRIKE: Fact$_0$(THIEF) = *strike*, etc.

Given the regular translation equivalents THIEF ≡ VOR, HURRICANE ≡ URAGAN, etc. (found in a bilingual index), the equivalences between the corresponding values of their Fact$_0$ are obtained automatically.

Example 2: The Polish verb OBOWIĄZYWAĆ 'oblige'

Sentence (4a) presents a sign seen in the hall of a Warsaw building, the Russian translation of which is given in (4b):

(4) a.

Mieszkańców	*budynku*	*obowiązuje*	*cisza*	*nocna*
tenant-PL.ACC	building-SG.GÉN	oblige-IND.PRES.3SG	silence-SG.NOM	nocturnal

lit. 'Nocturnal silence obliges tenants of [the] building'.

b.

Žiteli	*doma*	*objazany*	*noč′ju*	*sobljudat′*	*tišinu*
tenant-PL.NOM	building-SG.GÉN	are.obliged	at.night	observe	silence-SG.ACC

lit. 'Tenants of [the] building are.obliged at.night to.observe silence'.

From the viewpoint of translation of (4a) into (4b), (4a) presents two difficulties: the translation of OBOWJĄZYWAĆ 'oblige' and that of the expression *cisza nocna* 'nocturnal silence'.

The first difficulty is the specific government of the Polish verb, because of which it does not have a direct Russian (or English, for that matter) equivalent: **tišina objazyvaet* . . . *'the silence obliges . . .'. To ensure a correct rendering, a Polish ECD must contain, for the verb OBOWIĄZYWAĆ, the following indication:

X obowiązuje Y-a ⇔ *Y is obliged to Real$_1$(X)*-**II**→*X*

A Russian ECD has, under TIŠINA, Real$_1$(TIŠINA 'silence') = *sobljudat′* 'observe', which allows for the construction of sentence (4b)'s initial part: *Žiteli doma objazany sobljudat′ tišinu*. (The English ECD has, under SILENCE, Real$_1$(SILENCE) = //*be quiet*.)

The second difficulty concerns a particularity of the Russian lexicon: the noun TIŠINA 'silence' has the collocate *nočnaja* 'nocturnal', which corresponds exactly to the Polish adjective *nocna* 'nocturnal', however, **sobljudat′ nočnuju* *tišinu* is not a proper way of saying this; you have to put it as follows: *noč′ju sobljudat′ tišinu*. To resolve this difficulty it is sufficient to indicate, under TIŠINA 'silence', in the subarticle of Real$_1$(TIŠINA) = SOBLJUDAT′, that the temporal or locative modifier of the noun TIŠINA must be syntactically transferred to the verb SOBLJUDAT′:

*SOBLJUDAT′-**II**→TIŠINA-**ATTR**→$\Xi_{(«temp»/«loc»)}$ ⇒

$\Xi_{(«temp»/«loc»)}$←**ATTR**-SOBLJUDAT′-**II**→TIŠINA

(This particular rule corresponds to a general rule of deep-syntactic paraphrasing: Mel'čuk 1992: 50, Rule nº 19; for instance: *They launch **regular***←**ATTR**-*attacks*. ≡ *They **regularly***←**ATTR**-*launch attacks*.) In our case, the result is *soubljudat′*-**ATTR**→*noč′ju tišinu*.

Example 3: The French noun APPOINT 'exact sum paid by X to Y for Z such that Y does not have to give X any change' (= *exact change*)

A sign on a French bus shown in (5a) has a possible Russian translation in (5b):

(5) a. *Merci de faire l'appoint* lit. 'Thank.you for doing the exact.change'.
b. *Platite za proezd bez sdači* lit. 'Pay for transportation without change'.

This equivalent can be produced, using a pair of dictionaries of the ECD type, in five steps.

- *Merci de Y* is a pragmatically constrained lexeme that must be described in a French ECD as a non-standard LF under PRIER 'ask':
«Authorities ask you to Y» : *Merci* [de V('Y')$_{\mathrm{INF}}$] [**on a sign**]
- PRIER has a regular Russian equivalent PROSIT′ 'ask' (in a bilingual index). Under PROSIT′, the Russian ECD has the above non-standard LF:
«Authorities ask you to Y» : V('Y')$_{\mathrm{IMPER.2PL}}$ [**on a sign**]
- *Faire* in (5a) is described in the French ECD as Real$_1$ of APPOINT:
Real$_1$(APPOINT) : *faire* [*l'*~]
- APPOINT is translated (in the bilingual index) as PLATA BEZ SDAČI lit. 'sum paid by X to Y for Z such that Y does not have to give X any change'.
- Real$_1$(PLATA 'the sum paid') : //*platit′* 'pay'

These five steps lead to *Platite bez sdači* lit. 'Pay without change'. But Russian also requires the indication of the thing paid for: *platit′ za čto?* 'pay for what?' – *za proezd* 'for transportation'. This indication can be extracted from general knowledge about the situation in which the relevant phrase is used: if the sign is placed in a public transportation vehicle, you have to add *za proezd*; if it is hung on a ticket office, *za bilet* 'for ticket' is a must; if it is over the counter of a diner, it will read *za obed* 'for lunch'.

There is another way, perhaps even simpler, to establish the equivalence in question; namely, *faire l'appoint* can be described as a non-standard LF of PAYER:

```
    by giving to Y the exact sum due, so that
    Y does not have to give the change to X      : //faire l’appoint
```

The corresponding non-standard LF in Russian is given under PLATIT′ ‘payer’:

PLATIT′ ‘payer’

```
    by giving to Y the exact sum due, so that
    Y does not have to give the change to X      : bez sdači lit.
                                                   ‘without change’
```

The equivalence is then obtained in one step. Nevertheless, I wanted to present multiple paths that could lead to the same result.

Because of their huge numbers in every language and their importance for automatic processing of texts, collocations are increasingly becoming explored in the perspective of automatic discovery and extraction (see, for instance, Ferraro et al. 2012, with additional references).

5 Conclusions

Here are the five most important points of the present article:

1. Phrasemes constitute a significant part of the lexical stock of any language; therefore, they have to be presented in a formal dictionary of **L** (of the ECD type) in a systematic way.
2. A dictionary of the ECD type is the key for the automatic production of high quality texts.
3. Such a dictionary must reserve a place of honor for collocations described in terms of lexical functions.
4. LFs must be exploited in two major respects: for lexical selection and for paraphrasing.
5. A paraphrasing system must be part of any reliable NLP system.

Based on our today’s knowledge about phraseology, it is possible to sketch the following directions for future research:

- Massive inventarization of clichés (including pragmatemes) for major languages of the world. As of today, this type of phrasemes is seriously understudied.
- Deeper theoretical analysis of clichés, first of all – their finer typology and their conceptual characterization.
- Elaboration of methodologies and techniques for a better representation of all types of phrasemes in general dictionaries.

Observatoire de linguistique Sens – Texte, Université de Montréal

Acknowledgments

The text of the paper has been read and criticized by M. Alonso Ramos, D. Beck, L. Iordanskaja, J. Milićević, A. Polguère, Z. Saloni, and L. Wanner, whom I thank from the depth of my heart. I also profited a lot from long and heated discussions concerning phrasemes with S. Kahane and A. Polguère. And the two anonymous reviewers for the *Yearbook of Phraseology* proposed several significant improvements. I did my best to account for their remarks and suggestions; but of course I *bear* [= Oper$_1$(RESPONSIBILITY)] the *full* [= Magn(RESPONSIBILITY)] *responsibility* for any mistake that survived their scrutiny.

References

Alonso Ramos, Margarita. 2004a. *DiCE = Diccionario de Colocaciones del Español*. See: http://www.dicesp.com/paginas [accessed January 7, 2012].

Alonso Ramos, Margarita. 2004b. *Las construcciones con verbo de apoyo*. Madrid: Visor Libros.

Alonso Ramos, Margarita. 2004c. Elaboración del *Diccionario de colocaciones del español*. In P. Bataner & J. de Cesaris (eds.), *De lexicografía. Actes del I Symposium internacional de lexicografía*, Barcelona: IULA, 149–162.

Alonso Ramos, Margarita. 2005. Semantic description of collocations in a lexical database. In: Ferenc Kiefer, Gábor Kiss & Júlia Pajzs (eds.), *Papers in computational lexicography COMPLEX 2005*, 17–27.

Alonso Ramos, Margarita. 2010. "No importa si la llamas o no colocación, descríbela". In Carmen Mellado Blanco, Patricia Buján Otero, Claudia Herrero Kaczmarek, Nely Iglesias & Ana Mansilla Pérez (eds.), *La fraseografía del S.XXI. Nuevas propuestas para el español y el alemán,* 55–80. Berlin: Frank & Timme.

Álvarez de la Granja, María (ed.). 2008. *Fixed expressions in cross-linguistic perspective: A multilingual and multidisciplinary approach*. Hamburg: Verlag Dr. Kovac.

Anscombre, Jean-Claude & Salah Mejri (eds.). 2011. *Le figement linguistique: La parole entravée*. Paris: Honoré Champion.

Apresjan, Jurij & Leonid Cinman. 1998. Perifrazirovanie na kompʹjutere [Paraphrasing with computer]. *Semiotika i informatika* 36. 177–202.

Apresjan, Jurij & Leonid Cinman. 2002. Formalʹnaja modelʹ perefrazirovanija predloženij dlja sistem pererabotki tekstov na estestvennyx jazykax [A formal model of sentence paraphrasing for natural language text processing systems]. *Russkij jazyk v naučnom osveščenii*, No. 2 [= 4]. 102–146.

Bally, Charles. 1951 [1909]. *Traité de stylistique française*. Genève: Georg et Cie and Paris: Klincksieck.

Beck, David & Igor Mel'čuk. 2011. Morphological phrasemes and Totonacan verbal morphology. *Linguistics* 49(1). 175–228.

Becker, Joseph. 1975. The phrasal lexicon. In *Proceedings of the Workshop on Theoretical Issues in NLP*, 70–73. Cambridge, MA: MIT Press. See also: http://acl.ldc.upenn.edu/T/T75/T75–2013.pdf [accessed January 7, 2012.]

Bogusławski, Andrzej & Jan Wawrzyńczyk. 1993. *Polszczyzna, jaką znamy: Nowa sonda słownikowa* [Polish as we know it: A new lexicographic probe]. Warszawa: Uniwersytet Warszawski.

Bogusławski, Andrzej & Magdalena Danielewiczowa. 2005. *Verba polona abscondita. Sonda słownikowa III* [Concealed Polish words. Lexicographic probe III]. Warszawa: Uniwersytet Warszawski.

Bosredon, Bernard. 2011. Dénominations monoréférentielles, figement et signalétique. In Jean-Claude Anscombre & Salah Mejri (eds.), *Le figement linguistique: La parole entravée*, 155–169. Paris: Honoré Champion.

Burger, Harald, Dmitrij Dobrovol′skij, Peter Kuhn & Neal Norrrick (eds.). 2007. *Phraseology: An international handbook of contemporary research: Vols. 1–2*. Berlin: de Gruyter.

Cowie, Antony (ed.). 1998. *Phraseology: Theory, analysis, and applications*. Oxford: Clarendon Press.

Cowie, Anthony, Ronald Makin & Isabel McCaig. 1993. *Oxford dictionary of English idioms*. Oxford: Oxford University Press.

Everaert, Martin, Erik-Jan van der Linden, André Schenk & Rob Schreuder (eds.). 1995. *Idioms: Structural and psychological perspectives*. Hillsdale, N.J. & Hove, UK: Lawrence Erlbaum.

Ferraro, Gabriela, Rogelio Nazar, Margarita Alonso Ramos & Leo Wanner. 2012. Towards advanced collocation error correction in Spanish learner corpora. *Language Resources and Evaluation.* [to appear]

Grossmann, Francis & Agnès Tutin (eds.). 2003. *Les collocations: Analyse et traitement*. [= *Travaux et recherches en linguistique appliquée, Série E: Lexicologie et lexicographie*, n$^{\underline{o}}$ 1]. Amsterdam: De Werelt.

Iordanskaja, Lidija, Myunghee Kim & Alain Polguère. 1996. Some procedural problems in the implementation of lexical functions for text generation. In Leo Wanner (ed.), *Lexical functions in lexicography and natural language processing*, 279–297. Amsterdam and Philadelphia: John Benjamins.

Kahane, Sylvain. 2003a. The Meaning–Text Theory. In Ágel, Vilmos, Ludwig M. Eichinger, Hans Werner Eroms, Peter Hellwig, Hans Jürgen Heringer & Henning Lobin (eds.), *Dependency and valency: An international handbook of contemporary research*, vol. 1, 546–570. Berlin and New York: W. de Gruyter.

Kahane, Sylvain. 2003b. Sur le lien entre la définition lexicographique et les fonctions lexicales: une blessure profonde dans le DEC. In Francis Grossman & Agnès Tutin (eds.), *Les collocations: Analyse et traitement* [= *Travaux et recherches en linguistique appliquée, Série E: Lexicologie et lexicographie*, n$^{\underline{o}}$ 1], 61–73. Amsterdam: De Werelt.

Kahane, Sylvain & Alain Polguère. 2001. Formal foundations of lexical functions. In *Proceedings of* "*COLLOCATION: Computational Extraction, Analysis and Exploitation*"*, 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics*, 8–15. Toulouse.

Lareau, François & Leo Wanner. 2007. Towards a generic multilingual dependency grammar for text generation. In Tracy Holloway King & Emily Bender (eds.), *Proceedings of the Grammar Engineering Across Frameworks 2007 Workshop. CSLI Publications*, Stanford, CA: Stanford University. See also: http://www-csli.stanford.edu/~thking/GEAF07.html [accessed January 7, 2012].

Lubensky, Sophia. 1995. *Russian-English dictionary of idioms*. New York: Random House.

Mel'čuk, Igor. 1974. *Opyt teorii lingvističeskix modelej Smysl – Tekst* [Outline of a Meaning-Text type linguistic models]. Moskva: Nauka. [Reprint: 1999.]

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

Mel'čuk, Igor. 1982a. Lexical functions in lexicographic description. In *Proceedings of the VIIIth Annual Meeting of the Berkeley Linguistics Society*, 427–444. Berkeley: UCB.

Mel’čuk, Igor. 1982b. *Towards a language of linguistics*. München: W. Fink Verlag.

Mel’čuk, Igor. 1988a. *Dependency syntax: Theory and practice*. Albany, N.Y.: State University of New York Press.

Mel’čuk, Igor. 1988b. Semantic description of lexical units in an Explanatory Combinatorial Dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, 1(3), 165–188.

Mel’čuk, Igor. 1988c. Paraphrase et lexique dans la théorie linguistique Sens-Texte. In Gabriel Bès & Catherine Fuchs (eds.), *Lexique et paraphrase* [= *Lexique* 6], 13–54. Lille: Presses Universitaires de Lille.

Mel’čuk, Igor. 1992. Paraphrase et lexique: Vingt ans après. In Igor Mel’čuk avec Nadia Arbatchewsky-Jumarie & Suzanne Mantha. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-semantiques III*, 9–58. Montréal: Les Presses of l’Université de Montréal.

Mel’čuk, Igor. 1995. Phrasemes in language and phraseology in linguistics. In Martin Everaert, Erik-Jan van der Linden, André Schenk & Rob Schreuder (eds.), *Idioms: Structural and psychological perspectives*, 167–232. Hillsdale, N.J. and Hove, UK: Lawrence Erlbaum.

Mel’čuk, Igor. 1996. Lexical functions: A tool for the description of lexical relations in a lexicon. In Leo Wanner (ed.), *Lexical functions in lexicography and natural language processing*, 37–102. Amsterdam and Philadelphia: John Benjamins.

Mel’čuk, Igor. 1997. *Vers une linguistique Sens-Texte. Leçon inaugurale*. Paris: Collège de France.

Mel’čuk, Igor. 2001. *Communicative organization in natural language: The semantic-communicative structure of sentences*. Amsterdam and Philadelphia: John Benjamins.

Mel’čuk, Igor. 2003a. Collocations dans le dictionnaire. In Thomás Szende (ed.), *Les écarts culturels dans les dictionnaires bilingues*, 19–64. Paris: Honoré Champion.

Mel’čuk, Igor. 2003b. Les collocations: definition, rôle et utilité. In Francis Grossmann & Agnès Tutin (eds.), *Les collocations: Analyse et traitement.* [= *Travaux et recherches en linguistique appliquée, Série E: Lexicologie et lexicographie*, n 1], 23–31. Amsterdam: De Werelt.

Mel’čuk, Igor. 2004. Verbes supports sans peine. *Lingvisticæ Investigationes*, 27(2), 203–217.

Mel’čuk, Igor. 2006a. Explanatory Combinatorial Dictionary. In Giandomenico Sica (ed.), *Open problems in linguistics and lexicography*, 225–355. Monza (Italy): Polimetrica. See also http://www.polimetrica.com/?p=productsListandsWord=lexicography [accessed January 7, 2012].

Mel’čuk, Igor. 2006b. *Aspects of the theory of morphology*. Berlin and New York: Mouton de Gruyter.

Mel’čuk, Igor, André Clas & Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Bruxelles: Duculot.

Mel’čuk, Igor & Alain Polguère. 1991. Aspects of the implementation of the Meaning-Text model for English text generation. In Ian Lancashire (ed.), *Research in Humanities* 1, 204–215. Oxford: Clarendon Press.

Mel’čuk, Igor & Alain Polguère. 2007. *Lexique actif du français: L’apprentissage du vocabulaire fondé sur 20 000 dérivations semantiques and collocations du français*. Bruxelles: De Boeck.

Mel’čuk, Igor & Leo Wanner. 1996. Lexical function and lexical inheritance for emotion lexemes in German. In Leo Wanner (ed.), *Lexical functions in lexicography and natural language processing*, 209–278. Amsterdam and Philadelphia: John Benjamins.

Mel’čuk, Igor & Leo Wanner. 2006. Syntactic mismatches in machine translation. *Machine Translation* 21, 81–138.

Mel’čuk, Igor & Alexander Zholkovsky. 1984. *Explanatory combinatorial dictionary of modern Russian*, Wien: Wiener Slawistischer Almanach.

Mel’čuk, Igor avec Nadia Arbatchewsky-Jumarie, Léo Elnitsky, Lidija Iordanskaja & Adèle Lessard. 1984. *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-semantiques* I Montréal: Les Presses de l’Université de Montréal.

Mel’čuk, Igor avec Nadia Arbatchewsky-Jumarie, Louise Dagenais, Léo Elnitsky, Lidija Iordanskaja, Marie-Noëlle Lefebvre & Suzanne Mantha. 1988. *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-semantiques* II. Montréal: Les Presses de l’Université de Montréal.

Mel'čuk, Igor avec Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja & Suzanne Mantha. 1992. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques* III. Montréal: Les Presses de l'Université de Montréal.

Mel'čuk, Igor avec Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Suzanne Mantha & Alain Polguère. 1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques* IV. Montréal: Les Presses de l'Université de Montréal.

Milićević, Jasmina. 2007. *La paraphrase: Modélisation de la paraphrase langagière*. Bern etc.: Peter Lang.

Nattinger, James. 1980. A lexical phrase grammar for ESL. *TESOL Quarterly* 14(3), 337–344.

Rey, Alain & Sophie Chantreau. 1993. *Dictionnaire des expressions et locutions*. Paris: Dictionnaires Le ROBERT.

Vincze, Orsolya, Estela Mosqueira & Margarita Alonso Ramos. 2011. An online collocation dictionary of Spanish. In Igor Boguslavsky & Leo Wanner (eds.), *Proceedings of the 5th International Conference on Meaning-Text Theory, Barcelona, September 8–9*, 2011, 275–286. See http://www.dicesp.com/app/webroot/files/file/MTT%202011%20 Vincze%20et%20al_.pdf [accessed January 7, 2012.]

Wanner, Leo (ed.). 1996. *Lexical functions in lexicography and natural language processing*. Amsterdam and Philadelphia: John Benjamins.

Weinreich, Uriel. 1969. Problems in the analysis of idioms. In Jaan Puhvel (ed.), *Substance and structure of language*, 23–81. Berkeley and Los Angeles, CA: University of California Press. Reprinted in: Uriel Weinreich, *On semantics*, 1980, 208–264. Philadelphia: University of Pennsylvania Press.

Žolkovskij, Aleksandr & Igor Melʹčuk. 1967. O semantičeskom sinteze [On semantic synthesis]. *Problemy kibernetiki* 19, 177–238.