Semantics and the Lexicon in Modern Linguistics

Igor Mel'čuk

In the article, main features of the Meaning–Text linguistic theory are outlined. The specific and very important role of the formalized lexicon is stressed. Samples of level-specific representations of the meaning under development are given with short explanations.

1 THEORETICAL FRAMEWORK

The presentation that follows is based on the Meaning-Text linguistic theory [= MTT]. Without entering into the details, I will simply indicate that this theory puts forward the idea of a formalized model of natural language—a Meaning-Text Model [= MTM], which is a system of rules able to 'mimic' the linguistic behavior of humans. More specifically, an MTM is aimed at performing the transition from what is loosely called meanings (roughly, any information, or content, that a speaker may be willing to transmit by means of his language) and texts (physical manifestations of speech). A core component of an MTM, where the biggest part of data about specific language is stored, is a semantically-oriented formalized lexicon; in the MTT, such a lexicon is called an Explanatory Combinatorial Dictionary [= ECD].

An ECD-type lexicon must, and I think in the nearest future will, be one of the main components of any linguistic description. In conjunction with a formal grammar of the language (syntax + morphology), it ensures meaning-to-text and text-to-meaning transitions. In other words, as the first step, it allows the MTM to establish correspondences between a given Sem(antic) R(epresentation) and all D(eep-)Synt(actic)Rs that correspond to it. Then, the MTM goes from a given DSyntR to all (alternative) phonetic strings that, according to the speakers, may implement it as the signifiers of (more or less synonymous) real sentences. Roughly, the vocation of an MTM—and of an ECD of course—is as follows:

From a Sem-**network** —
to all corresponding Deep-Synt- and Surface-Synt-**trees** —
to all corresponding Deep-Morph- and Surface-Morph-**strings** —
to all corresponding Phonemic strings!

In more precise terms, in the MTT, sentence *representation* at a particular level is a set of formal objects called *structures*, each of which is responsible for a particular aspect of sentence organization at this level. The set of sentence representations of all levels is as follows (starting from MEANING, i.e. Semantic Representation, and going to TEXTS, i.e. Surface-Phonological [= phonetic] Representation):

Sem(antic) Representation	=	<Sem-S(tructure); Sem-Comm(unicative) S; Sem-Rhetorical S>
D(eep)-Synt(actic) Representation	=	<DSyntS; DSynt-CommS; DSynt-Anaph(orical) S; DSynt-Pros(odic) S>
S(urface)-Synt(actic) Representation	=	<SSyntS; SSynt-CommS; SSynt-AnaphS; SSynt-ProsS>
DMorph(ological) Representation	=	<DMorphS; DMorph-ProsS>
SMorph Representation	=	<SMorphS; SMorph-ProsS>
DPhon(ological) Representation	=	<DPhonS; DPhon-ProsS>
SPhon(ological) Representation	=	<SPhonS; SPhon-ProsS>

The role of an ECD in the passage $\{SemR_i\} \Leftrightarrow \{SPhonR_j\}$ is crucial: it is in an ECD that the rules of the model find all the information which is associated with individual lexical units and which is necessary for the determination of well-formed configurations of linguistic signs that constitute actual sentences.

2 THE EXPLANATORY-COMBINATORIAL DICTIONARY

Given the central position of the ECD in an MTM, I will briefly characterize this dictionary. Its main feature is that it is PARAPHRASE-BASED, that is, (quasi-synonymous) paraphrases constitute the main target and the main research tool for an ECD; cf. the sample set of such paraphrases in **5**. This means that the ECD's foundations are SEMANTIC. I can state the following six major properties of an ECD that set it aside from many other dictionaries of linguistic orientation:

- An ECD is elaborated within a coherent linguistic theory: the Meaning-Text theory, featuring a developed semantic and syntactic modules, with a strong emphasis on the lexicon.

- An ECD is formally linked to a grammar (syntax + morphology); both are 'tuned' to each other, so that the lexicon and the grammar are in complete logical agreement.
- An ECD is consistently geared to production: it is a synthesis (= active) dictionary.
- An ECD is centered around restricted cooccurrence, both syntactic and lexical.
- An ECD is an integral dictionary: it includes all of the information that is lexically-related and could be needed for successful text-synthesis.
- An ECD is a formalized dictionary (= a lexical database).

3 LEXICAL UNIT

A unit of description in an ECD is a *lexical unit* — a word (= lexeme) or a set phrase (= full phraseme or quasi-phraseme, see Mel'čuk 1995, 1996) taken in one well-defined sense.

> An extremely fine sense discrimination is one of the slogans of the ECD.

4 THE STRUCTURE OF AN ECD ENTRY

The ECD entry for a lexical unit L—lexeme or phraseme—has three main zones.

1) The semantic zone: the definition of L (= a SemR of L), which is based on a propositional form with variables for semantic actants of L and constitutes a strict decomposition of its meaning. For instance, the verb [*to*] HELP (in one of several senses; the symbol "||" separates the presuppositional part of the definition to the left of it from the assertional part to the right of it):

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 facilitate for Y doing Z'.

X's revulsion for Y : 'X perceiving Y, || X's (strong) negative emotion about Y which is similar to what people normally experience when they perceive something that makes them sick and such that it causes that X wants to avoid any contact with Y'.

X is a challenge for Y : 'Y having to do X, || X is difficult but interesting for Y, which causes that Y wants to do X'.

These definitions are written in standardized English for the convenience of the reader and in order to facilitate the task of their authors and critics: the linguistic intuition of a speaker permits better judgments when applied to such linguistic expressions, even if they are in 'processed' English. For formal treatment, each definition has a corresponding representation in the form of a semantic network, see below.

2) The SYNTACTIC zone: the Government Pattern (= a subcategorization frame) of L, which specifies, for each Sem-actant of L, the corresponding DSynt-actant and lists all surface means of expressing it in the text as a function of L. Cf. the Government Pattern [= GP] for the verb (*to*) HELP:

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

1) $C_{III.1}$: 'X being directly involved in Z' [= 'X doing Z'] [C stands for *column*]

2) $C_{III.2}$: 'X not being directly involved in Z' [= 'X does not do Z himself, but provides some resources to Y'] **or** H. is in the passive

3) $C_{III.4}$: Z = 'move $PREP_{dir}$ N'

Frederique helped the old gentleman finish his preparations <helped the boy to finish his studies with her generous financial assistance, helped Jack out of his coat, helped Jack up the stairs by a kick in the bottom /by pushing him hard>.

Through Government Patterns,
SemRs of lexical units link to syntax.

3) The LEXICAL zone: Lexical Functions [= LF] of L, which present, in a systematic and formal way, the whole of semantic derivation (paradigmatic lexical functions) and of restricted lexical cooccurrence of L—i.e., all of its collocations (syntagmatic lexical functions). Paradigmatic LFs correspond to

derivational relations well-known in linguistics: synonymy, antonymy, conversion, nominalization, agent/patient noun, relative adjective, etc. A syntagmatic LF **f** is, roughly speaking, a very general and abstract meaning which can be expressed in a large variety of ways depending on L, which is the argument of **f**. For instance:

Magn(L): intensifier of L, i.e. a modifier that expresses a high degree of what is designated by L; ≈ 'very', 'very much', 'completely'

Oper$_1$(L): support verb of L, i.e. a semantically empty verb that takes the first DSynt-actant of X as its subject and L itself as its main object'; ≈ 'do', 'make', 'have'

Real$_i$(L): verb of realization for L, i.e. a semantically full verb that means '[the *i*-th DSynt-actant of L] does with L what this actant is supposed to do with X'; ≈ 'succeed', 'use', 'accomplish'

English

Magn(*naked*)	= *stark*	**Oper$_1$**(*sovereignty*)	= *have* [~]
Magn(*thin*)	= *as a rake*	**Oper$_1$** (*cry*)	= *let out* [ART ~]
Magn(*patience*)	= *infinite*	**Oper$_1$** (*whack*)	= *fetch* [*a* ~]
Magn(*rely*)	= *heavily*	**Oper$_1$** (*support*)	= *lend* [~]

Real$_2$ (*joke*)	= *get* [ART ~]
Real$_2$ (*demands*)	= *meet* [~]
Real$_2$ (*exam*)	= *pass* [ART ~]
Real$_2$ (*hint*)	= *take* [ART ~]

Spanish

Magn(*loco*)	= *como una cabra*	**Oper$_1$** (*siesta*)	= *echar* [ART ~]
Magn(*tráfico*)	= *denso*	**Oper$_1$** (*cuestión*)	= *plantear* [ART ~]
Magn(*silencio*)	= *profundo*	**Oper$_1$** (*juramento*)	= *prestar* [~]
Magn(*comer*)	= *a dos carrillos*	**Oper$_1$** (*resistencia*)	= *poner* [~]

Real$_1$ (*tesis*)	= *leer* [ART ~]
Real$_1$ (*botón*)	= *pulsar* [ART ~]
Real$_2$ (*condición*)	= *cumplir* [ART ~]
Real$_3$ (*orden*)	= *ejecutar* [ART ~]

LFs of a lexical unit L consistently link with the definition of L, so that if, e.g., L has a **Real$_i$**, it must have in its definition the corresponding semantic component: 'such that X is supposed to ...', etc.

5 A SAMPLE SET OF APPROXIMATE PARAPHRASES

Consider the sentence (1):

(1) *What has been discovered lends strong support to the view that the progress which lead to the most advanced Pre-Columbian society may have occurred much earlier than was previously hypothesized, in the words of Richard Hansen.*

The meaning expressed in (1) can be expressed as well by more than a million and a half other English sentences which can be constructed form the set of near-synonymous expressions given below:

This find What has been {found / discovered} The {things / objects} {found / discovered}	clearly {shows / indicates} {give(s) / suppli(es)} {clear indication / convincing argumen convincingly demonstrate(s) lend(s) strong support to the view	that {the achievements / the progress / the developments / the advances}
$7 \times 8 \times 4 = 224$		

{which / that} {produced / created / lead to}	the most sophisticated the most advanced the most developed	Pre-Columbian Society
$6 \times 3 \times 1 = 18$		

{may have / have probably}	occurred taken place happened	much earlier than much before what long time before the date that
$2 \times 3 \times 3 = 18$		

was	before previously	assumed thought hypothesized believed	(as) said Richard Hansen according to R. Hansen in the words of R. Hansen
$1 \times 2 \times 4 \times 3 = 24$			

$224 \times 18 \times 18 \times 24 = 1\ 741\ 824$ paraphrases are available for sentence (1)

Such paraphrastic sets underlie SemRs used by the Meaning-Text theory and the Explanatory Combinatorial Dictionary: when discussing a specific SemR, in particular a lexicographic definition, the researcher uses paraphrases as his arguments and his source of linguistic insights. On the other hand, a SemR underlies such a paraphrastic set in a different sense: all the paraphrases in the set must be obtainable from this SemR by the rules of the corresponding Meaning-Text Model.

6 A SAMPLE SEMANTIC REPRESENTATION

See Figure 1.

Explanations

- Semantic Structure (Predicates)

  'time-of→X': moment in which X takes place ('the time of saying is before now' ⇔ *said*)

  'X←is.challenge-for→Y' ('something that is part of excavating the site by $persons_i$ is a challenge for $persons_i$')

  'X←is.certain-of→Y' ('X says that Y must have taken place')

  'X←is.of.ethnicity→Y' ('people of ethnicity Maya' = 'Mayans')

  'X who says something belongs to the group α': speaking of α, X can say *we*

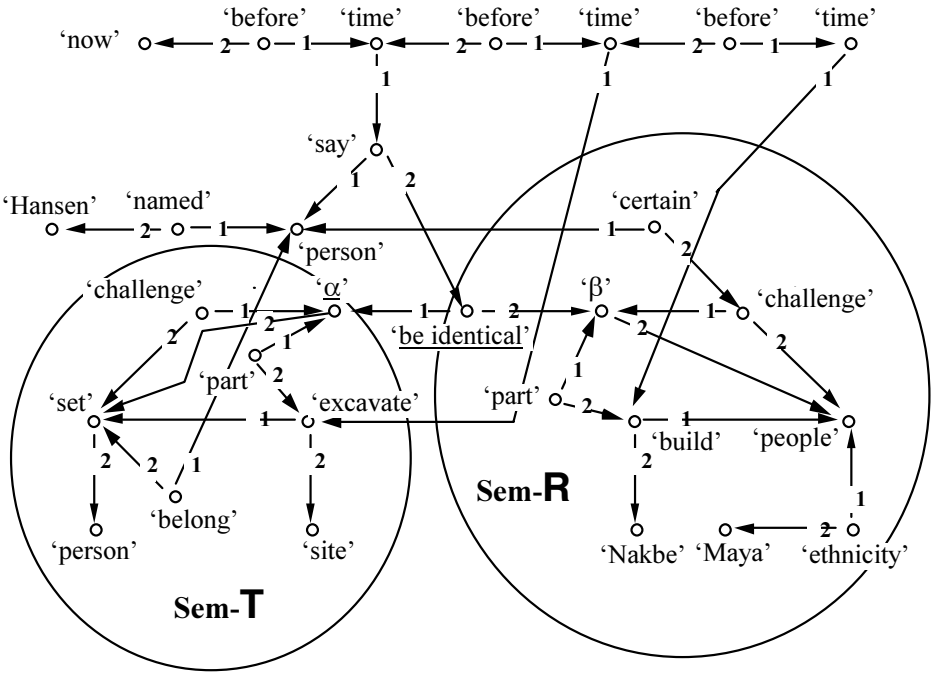


Figure 1.

'α' and 'β' are abbreviations for particular actions and states of $persons_i$; the actants of 'α' and 'β' are not shown in order to avoid cluttering the diagram with too many details

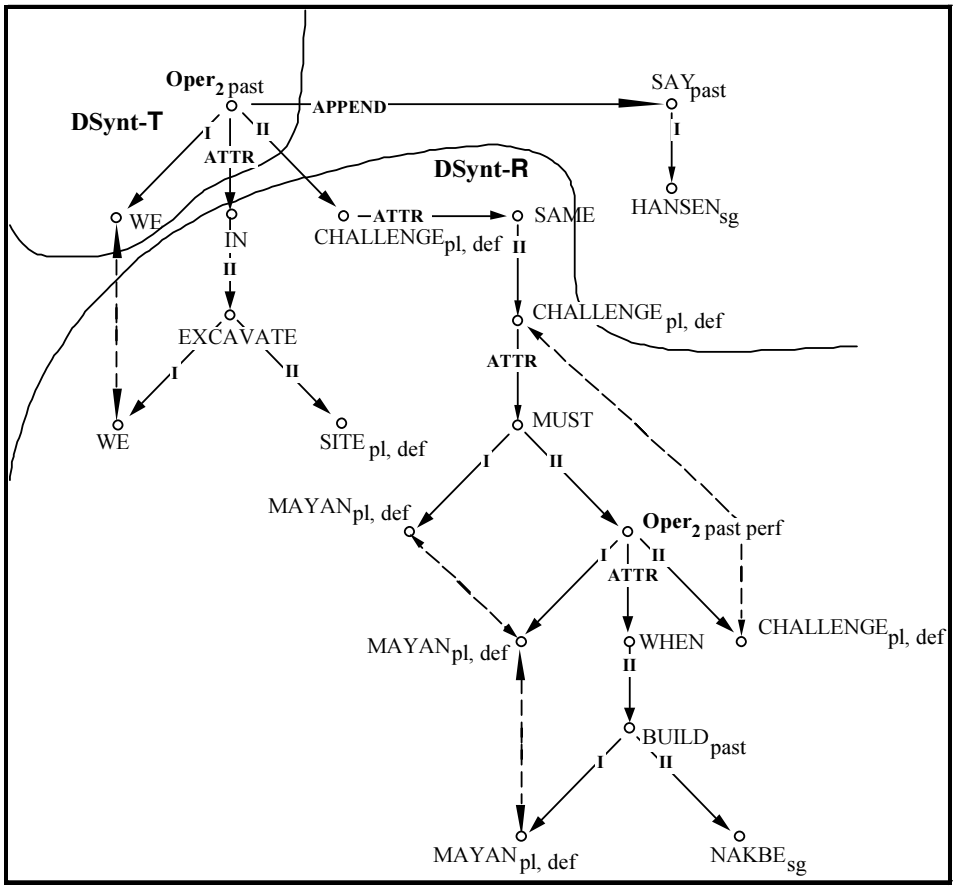
- Sem-Communicative Structure

The Semantic Theme Sem-**T** of the starting meaning is 'SOMETHING [= 'α'] faced by the $persons_i$ excavating the site and being a challenge for them'; 'α' is the Comm-dominant node of the Sem-**T**(underscored).

About this Sem-**T** it is asserted that this 'α' IS IDENTICAL to ...; 'be.identical' is the Comm-dominant node of the Sem-**R** (also underscored).

Semantemes that remain outside of the Sem-**T** and the Sem-**R** are Comm-Specifiers (= semaNteme configurations that, so to speak, set the scene for the main statement and characterize this scene form the viewpoint of when, where, in what way, with what purpose, etc. the situation in question has taken or is taking place).

Here are a few English sentences that can express the above SemR:



Fugure 2.

(2) *"We experienced the same challenges in excavating the site that the Mayans must have encountered when they built Nakbe," Hansen said.*

(3) *Hansen said that difficulties which had to be faced by him and by other archeologists excavating the site were similar to those which Mayans probably had met with when erecting Nakbe*.

(4) *The problems into which had run the archeologists during the excavation of the site were, according to Hansen, very much like those which Mayans could have had experienced while building Nakbe*.

Note that the starting SemR is, so to speak, underspecified: it does not necessarily contain all the semantic details that are expressed in the sentences synthesized from it. In the transition "SemR ⇒ DSyntR" particular lexical units are brought in that can make the initial meaning more (or less) precise and

elaborate. Therefore, the paraphrases obtained from a SemR are not 100% synonymous: they can differ semantically, but in such a way that this is considered irrelevant in the given act of linguistic communication.

7 A SAMPLE DEEP-SYNTACTIC REPRESENTATION

I show here — see Figure 2 — the (partial) DSyntR of sentence (2), namely — its DSyntS and its DSynt-AnaphS (broken-line arrows show the coreference of some lexical nodes) plus a partial specification of its DSynt-CommS; the DSynt-ProsS is not presented at all.

Explanations

Possible values of the Lexical Function **Oper$_2$**(*challenge*) are *experience* [ART ~] and *encounter* [ART ~]. (**Oper$_2$** is similar to **Oper$_1$**: it is also a support verb of a lexical unit L, i.e. a semantically empty verb that takes the second DSynt-actant of L as its subject and L itself as its main object'; ≈ 'undergo', 'be implicated', 'receive'.)

The DSyntS is supplied with its own DSynt-Communicative Structure: *We experienced...* constitutes the DSynt-**T** and *...the same challenges in excavating the site that ...*, the DSynt-**R**; *Hansen said* is a parenthetical that is outside of the main DSynt-Communicative division of the sentence and therefore it can be linearly placed either after the **T** or after the **R**.

8 A SAMPLE LEXICAL ENTRY OF AN EXPLANATORY COMBINATORIAL DICTIONARY

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 to avoid any contact with Y.

Government Pattern

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

1) $C_{II.2}$: N does something that can be seen or felt [but not heard!]

*John's revulsion *at* <correct: *for*> *these shouts*

2) $C_{II.4}$: N denotes people

John's <his> revulsion against racism <against Mary's greed>; *John's <his> revulsion at such behavior <at the sight of sea food>*;
John's <his> revulsion for work <for all those killings; *for this melody*, *for/toward all those scoundrels>*;
John's <his> revulsion toward the government

Lexical Functions

Syn_i	: distaste
Syn_i	: repugnance; repulsion; disgust; loathing
$Anti_i$	: attraction
$Conv_{21}Anti_i$	: appeal
A_1	: revulsed
Adv_1	: in [~]
$MagnAdv_1$	: well up in [~]
Propt	: from [~]
$Able_2$	: revulsive
$Magn + Able_2$	: of utmost [~] | G = SCENE, SIGHT [G stands for the Syntactic Governor of the value of the LF in question]
$Qual_1$	: squeamish; overly sensitive
Magn	: violent < extreme < utmost
AntiMagn	: slight
$Oper_1$	: experience, feel [~ for/towards N]
$Conv_{21}Caus_2Oper_1$	: be driven [to ~]
$Magn + Oper_1$	: be filled [with ~ (about N)]
$Magn + Labor_2$	: fill [N with ~]
$Caus_2$	: revolt [N]

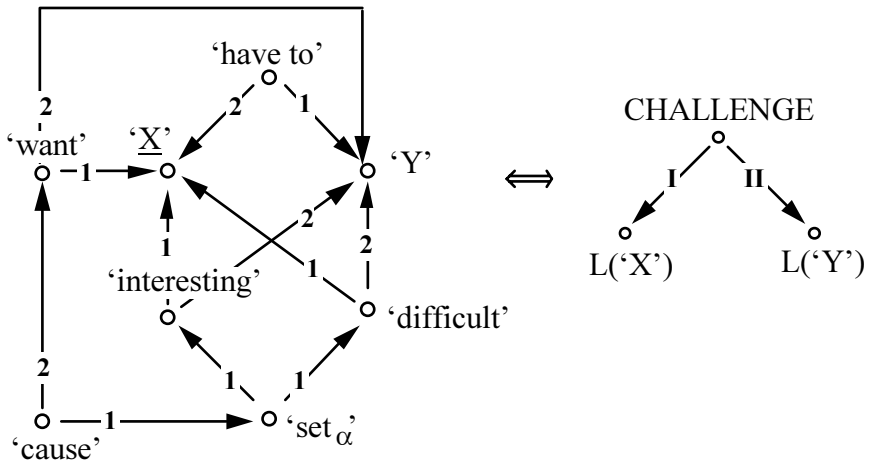
Adv_1Manif : with [~]

Examples

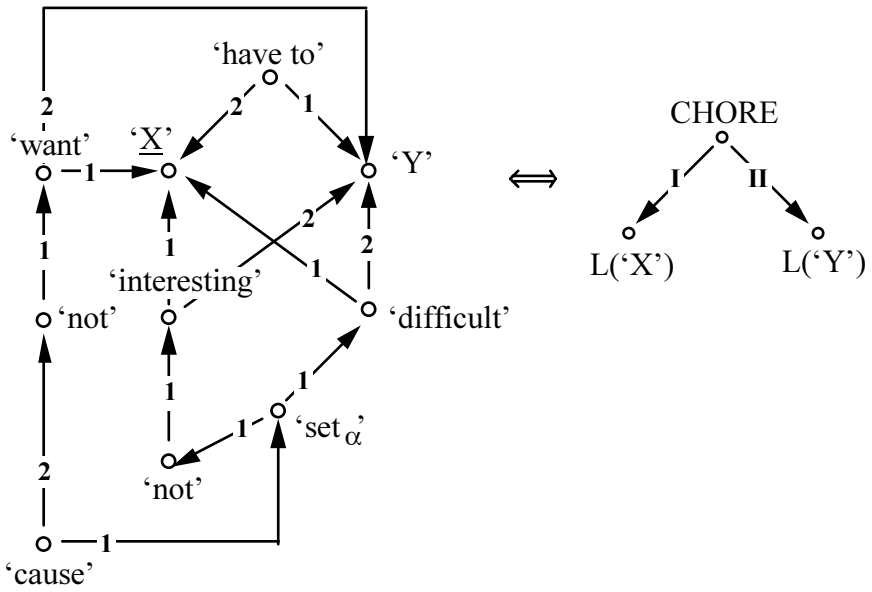
Any revulsion they might feel from fat-ass bastards they ran up against professionally was *ad hominem* and not *ad genus* [A. Lurie]. I felt no revulsion for her maternal fantasies, only a practical concern. She met his advances with revulsion ≈ She turned away in revulsion. It was a scene of utmost 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].

9 SAMPLE SEMANTIC RULES

Semantic rules are formal equivalents of lexicographic definitions, ECD-style. They constitute the Semantic Module of an MTM, that is, the module responsible for the correspondence $\{SemR_i\} \Leftrightarrow (DSyntR_j\}$. Here are the Sem-Rules for the lexemes CHALLENGE and CHORE:



This is a challenge for Alain; *They meet, encounter, face a challenge*.



This is a chore for Alain

In both cases, the rules describe the predicative expressions *X is a challenge/a chore for Y*, rather than the simple nouns CHALLENGE and CHORE.

One clearly sees the semantic differences between the two lexemes: *challenge X* is something interesting for Y, which makes Y want to do X; while *chore X* is something **not** interesting for Y, which makes Y **not** want to do X**.** CHALLENGE and CHORE are thus antonyms. A CHALLENGE is also something which is hard to accomplish, while a CHORE can be simply boring. Thus, these two lexemes are probably not exact antonyms, which I allowed myself to disregard in my approximate rules.

10 SEMANTIC DECOMPOSITIONS

Semantic decompositions are controled by paraphrastic equivalencies (= synonymy). Cf.:

X is sure that P : 'Having the belief «P has taken/is taking/will be taking place,» X is not disposed to admit that «P has not taken/is not taking/ will not be taking place»'

X doubts that P : 'Not having the belief «P has taken/is taking/ will be taking place,» X is disposed to admit that «P has not taken/is not taking/ will not be taking place»'

(5) *I believe that Arthur is in Montreal, but I am not sure.*

(6) *I am sure that Arthur is in Montreal,* $^{\#}$*but I don't believe that.*

[The symbol «$^{\#}$» indicates pragmatic or logical incorrectness.]

(7) *I believe that Arthur is in Montreal,* $^{\#}$*but I doubt it.*

(8) *I am sure that Arthur is in Montreal* = *I don't doubt that Arthur is in Montreal.*

(9) *I am not sure that Arthur is in Montreal* ≈ *I doubt that Arthur is in Montreal.*

MAIN REFERENCES

(An attempt at self-promotion)

Mel'čuk, I., *et al*. (1984, 1988, 1992, 1999) *Dictionnaire explicatif et combinatoire du français contemporain : Recherches lexico-sémantiques* I, II, III, IV. Montréal : Presses de l'Université de Montréal.

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

Mel'čuk, I., and A. Polguère (1987) A Formal Lexicon in the Meaning-Text Theory (or How to Do Lexica with Words). *Computational Linguistics*, 13 : 3/4, 261-275.

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

Mel'čuk, I. (1989) Semantic Primitives from the Viewpoint of the Meaning-Text Linguistic Theory. *Quaderni di semantica*, 10: 1, 65-102.

Mel'čuk, I., and R. Ilson (1989) English BAKE Revisited (BAKE-ing an ECD). *International Journal of Lexicography*, 2: 3, 325-345.

Mel'čuk, I., A. Clas, et A. Polguère (1995) *Introduction à la lexicologie explicative et combinatoire*. Bruxelles : Duculot.

Mel'čuk, I. (1995) *The Russian Language in the Meaning-Text Perspective*. Wiener Slawistischer Almanach / Škola "Jazyki russkoj kul´tury": Viena / Moscow.

Mel'čuk, I. (1995) Phrasemes in Language and Phraseology in Linguistics. In: M. Everaert, E.-J. van der Linden, A. Schenk and R. Schreuder (eds), *Idioms*. *Structural and Psychological Perspectives*, Hillsdale, N.J.—Hove: Lawrence Erlbaum Associates, 167-232.

Mel'čuk, I. (1996) Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In: L. Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/Philadelphia: Benjamins, 37-102.

Mel'čuk, I. (1997) *Vers une linguistique Sens-Texte. Leçon inaugurale.* Paris: Collège de France.

Polguère, A. (1997) Meaning-Text Semantic Networks as a Formal Language. In: L. Wanner (ed.), *Recent Trends in Meaning-Text Theory*, Amsterdam/Philadelphia: Benjamins, 1-24.

Steele, J., ed. (1990) *Meaning-Text Theory. Linguistics, Lexicography, and Implications*. Ottawa: Univ. of Ottawa Press.

Igor Mel'čuk is the author of the Meaning ⇔ Text theory of language understanding and translation. His main works are in the areas of morphology, computational lexicography and semantics, dependency syntax, and communicative structure of language. He works for Observatory of Meaning-Text Linguistics of the University of Montreal, Department of Linguistics and Translation, P.O. Box 6128 Centre-ville, Montreal H3C3J7, Canada. He can be reached at melcuk@ERE.UMontreal.CA.