Towards an explanatory-combinatorial dictionary of Japanese

François Lareau[1]
OLST, Département de linguistique et de traduction
Université de Montréal
C.P. 6128, succursale Centre-Ville
Montréal QC H3C 3J7
Canada
`francois.lareau@umontreal.ca`

Abstract

The Meaning-Text Theory is a formal, dependency-based linguistic theory in which the lexicon plays a central role. Within this theory, the vocabulary of a language is described by an Explanatory-Combinatorial Dictionary [ECD]. Unlike most existing dictionaries, an ECD is oriented towards language generation rather than analysis. This orientation makes it especially useful for language learners or computers that need to produce native-quality (i.e. highly idiomatic) texts.

An ECD of a language aims at describing exhaustively all three components of its lexical units: their meaning, their form and their combinatorics. While it is common for current dictionaries to describe words' meaning and form, their combinatorics is often left undescribed, so that the user has no information on how to actually use them in a sentence. An ECD, on the contrary, describes not only the syntactic and morphological behavior of words, but also their restricted lexical cooccurrence.

The description of the restricted lexical cooccurrence of words constitutes the heart of an ECD. It is done using a formal language known as lexical functions. Lexical functions offer an elegant and powerful way of encoding common semantico-syntactic patterns of lexical cooccurrence.

Until now, most of the work in explanatory-combinatorial lexicography has been done on French and Russian. The main goal of this paper is to show that it would be possible and useful to develop an ECD of Japanese, and to discuss potential problems inherent to this task and propose some leads towards a solution.

1 Introduction

Being a learner of Japanese, I have often been frustrated when trying to express myself in that language. Of course, this frustration is due to my huge lack of knowledge, but part of it comes from the absence of a good dictionary that could make up for my linguistic ignorance. Such a tool exists in other languages, Longman's *Language Activator* being a good example, but the only similar commercial dictionary that I have found in Montreal's bookstores for Japanese is Kodansha's *Dictionary of Japanese and English Idiomatic Equivalents*, which is a great piece of work but does not even come close to what is available in English. Is it that Montreal is too far from the civilized world, or is there really no such dictionary? Having been exposed for many years as a student to the *Explanatory-Combinatorial Dictionary of Contemporary French*, a formal dictionary oriented towards text production, my dream has been to find the equivalent in Japanese: an Explanatory-Combinatorial Dictionary [ECD] of Japanese. Since it does not exist yet, I will try to introduce what an ECD is, hoping that some more proficient speakers of Japanese than myself will have the great idea of developing one for that language.

Section 2 gives a general characterization of an ECD as well as a very brief introduction to the linguistic theory behind it. The formal language of lexical functions, used in an ECD to describe idioms, is then discussed in Section 3. Finally, as an illustration, I will examine in Section 4 the problem of Japanese numeral classifiers from a lexicographic point of view.

[1] Also affiliated to Lattice, CNRS unit 8094, University Paris 7.

2 The Meaning-Text Theory and the explanatory-combinatorial dictionary

The Meaning-Text Theory [MTT] (Mel'čuk 1981) is a dependency-based linguistic theory that was born in Moscow in the 60's. As most modern linguistic theories, it considers language as a mapping between a meaning and a form (in other words, a correspondence between meaning and text, thus the name of the theory). This correspondence is described by a functional model simulating the linguistic competence of human speakers. This correspondence being very complex to model, it is divided into intermediate modules. There are four major levels of representation of sentences, corresponding to widely accepted levels among the other current theories: semantics, syntax, morphology and phonology.

Rules of a Meaning-Text model describe signs of a language. Because the overwhelming majority of signs in any language are lexical, the lexicon plays a central role in MTT. Within this theory, the vocabulary of a language is described in an ECD (Mel'čuk 1995). An ECD of a language aims at describing exhaustively all three major components of its lexical units: their meaning, their form and their combinatorics. While it is common for dictionaries to describe words' meaning and form, their combinatorics is often left undescribed, so that the user has no information on how to actually use them in a sentence. An ECD, on the contrary, describes not only the syntactic and morphological behavior of words, but also their restricted lexical cooccurrence (i.e., idiomatic expressions controlled by the words). This, combined to the fact that unlike most existing dictionaries, ECDs are oriented towards language generation rather than analysis, makes them especially useful for language learners or text generation systems that need to produce native-quality (i.e. highly idiomatic) texts.

To date, there are two ECDs: one for French (Mel'čuk et al 1984, 1988, 1992, 1999) and one for Russian (Zholkovsky & Mel'čuk 1984). Besides these two dictionaries, there is also the *DiCo* lexical database. It is an electronic lexical database intended to be used by either machines or humans. It is therefore formalized, but in a way that is still easily readable for humans. From this database is derived a general public paper dictionary, the *Lexique Actif du Français*, which is similar to learner's dictionaries such as Longman's *Language Activator*. See (Polguère 2000) for information about the *DiCo* and *Lexique Actif du Français*. From the same database, it is also possible to derive in a totally automatic manner a fully formalized dictionary for direct use in computer systems (Lareau 2002). The linguistic information contained in the *DiCo* database is close to what one finds in an ECD, except for the definitions, which have been replaced by a much lighter semantic description (Milićević 1997, Polguère 2003), allowing for a broader coverage of the language.

In the *DiCo* as well as in ECDs, idiomatic expressions controlled by the lexical entries constitute the major part of the data. It is not surprising after all, since restricted lexical cooccurrence is a very recurrent phenomenon in human languages. It is very often that words do not combine freely. For instance, one says ***heavy*** *smoker*, ***heavy*** *rain* and ***heavy*** *penalty*, but ***high*** *salary* (**heavy salary*), ***true*** *love* (**heavy love*) and ***big*** *crush [on someone]* (**heavy crush*), even though all of these phrases bear exactly the same meaning 'intense / at a high level'. Because of their arbitrary nature, such phrases have to be described in the dictionary. In an ECD, this is done using a formal tool called *lexical functions*.

3 Lexical functions

Lexical functions are used to identify common links between lexemes (synonymy, antonymy, hyperonymy, etc.) or else recurrent semantico-syntactic patterns of idiomatic phrases called *collocations*. A collocation is a sequence of words where only one lexeme, the *base*, is chosen freely by the speaker, the choice of the other one(s), the *collocate*, being restricted by the base. For instance, among the examples given above, *heavy*, *high*, *true* and *big* are collocates whereas *smoker*, *rain*, *penalty*, *salary*, *love* and *crush* are bases. These collocates all bear the same meaning of intensification, so that one could say that the ratio *heavy* : *smoker* is equal to the ratio *high* : *salary*. The idea of lexical functions is to describe such patterns by a mathematical function ***f***(*x*) that would return, for each base, the appropriate collocate expressing the meaning identified by ***f***. Thus, we would have ***f***(SMOKER) = heavy, ***f***(SALARY) = high, etc.

Over the years, more than fifty basic meanings have been identified as frequently expressed by collocates in human languages and for each, a standard lexical function has been created. Due to space limitations, I cannot go through all the fifty-something standard lexical functions in this paper, but I will introduce a few selected ones. For a complete presentation of the whole set of functions and for an in-depth discussion, see (Mel'čuk 1995, Wanner 1996, Kahane & Polguère 2001). I provide at first examples in English rather than Japanese, assuming that all readers have sufficient knowledge of the former while some might not understand the latter. Examples in Japanese will follow.

3.1 Paradigmatic functions

The lexical functions that identify links between lexemes are called *paradigmatic functions*. The most important are the following:

- ***Syn***(*L*) returns synonyms of *L*. In the *DiCo*, the function ***QSyn*** is used for approximate synonyms: ***QSyn***(AGGREGATE$_{(V)}$) = combine, ***QSyn***(FEELING) = emotion, ***Syn***(PHONE$_{(V)}$) = telephone$_{[V]}$.
- ***Anti***(*L*) returns the antonyms of *L*. In the *DiCo*, the function ***QAnti*** is used for approximate antonyms: ***QAnti***(HAPPY) = sad, ***Anti***(HIGH) = low, ***Anti***(BEHAVE) = misbehave.
- ***S***$_i$(*L*) returns the typical name for the *i*-th actant of *L*: ***S***$_1$(BUY$_{(V)}$) = buyer, ***S***$_2$(BUY$_{(V)}$) = purchase$_{(N)}$, ***S***$_3$(BUY$_{(V)}$) = seller, ***S***$_4$(BUY$_{(V)}$) = price.

3.2 Syntagmatic functions

The lexical functions that identify collocations are called *syntagmatic functions*. Most lexical functions are of that type. Here are some of them:

- ***Sing***(*L*) returns nouns that bear the meaning ‘a unit of’. Such collocations are often used to make mass nouns, nouns that have no singular and the like, countable: ***Sing***(INFORMATION) = piece [of ~] (*Agent Wanner gave us a good piece of information*), ***Sing***(GARLIC) = clove [of ~] (*Crush two cloves of garlic*).
- ***Mult***(*L*) returns nouns that express the meaning ‘a group of’: ***Mult***(BEE) = swarm [of ~], ***Mult***(SHEEP) = flock [of ~], ***Mult***(PAPER) = pile [of ~].
- ***Magn***(*L*) returns collocates that mean ‘very’ / ‘intensely’: ***Magn***(RAIN$_{(N)}$) = heavy, ***Magn***(CARE$_{(N)}$) = extreme, ***Magn***(CHEAP) = dirt [~], ***Magn***(NOTHING) = at all, ***Magn***(SCARE$_{(V)}$) = [~] the hell out [of *Y*].
- ***Ver***(*L*) returns collocates that mean ‘as it should be’: ***Ver***(WRITING) = intelligible, ***Ver***(EXAMPLE) = clear.
- ***Oper***$_i$(*L*) and ***Func***$_i$(*L*) both return semantically empty verbs (known as “light verbs” or “support verbs”). The distinction between them is purely syntactic: ***Oper***$_i$(*L*) returns a verb which has the *i*-th actant of *L* as its subject and *L* as its object, while ***Func***$_i$(*L*) returns a converse verb which takes *L* as its subject and the *i*-th actant of *L* as its object[2]. *i*=0 when an ***Oper*** has a dummy subject or a ***Func*** has no object: ***Oper***$_1$(DISEASE) = have [a ~], ***Func***$_1$(DISEASE) = [~] affects [*X*], ***Oper***$_2$(BLOW$_{(N)}$) = receive [a ~], ***Func***$_2$(BLOW$_{(N)}$) = [~] falls [upon *Y*], ***Oper***$_0$(FOG) = there is [~], ***Func***$_0$(CHANGE$_{(N)}$) = [~] occurs.
- ***Real***$_i$(*L*) returns verbs that have the same syntax as ***Oper***$_i$(*L*) and bear the meaning ‘to do what you are supposed to do with’: ***Real***$_1$(GOAL) = achieve [a ~], ***Real***$_2$(ATTACK$_{(N)}$) = fall [to an ~].

3.3 Complex functions and configurations of functions

Lexical functions can combine together to form *complex functions*. For instance, ***Anti*** may combine with ***Magn*** to form a new function which returns collocates that mean the contrary (***Anti***) of intensity (***Magn***): ***AntiMagn***(INCREASE$_{(N)}$) = small, ***AntiMagn***(INCOME) = low. Some functions are used most of the time as components of complex functions. It is the case of the following six ones (phasal and causative verbs):

- ***Incep***(*L*), ***Cont***(*L*) and ***Fin***(*L*) return verbs that indicate respectively the beginning, the continuation and the end of a process: ***IncepOper***$_1$(DISEASE) = catch [a ~], ***ContFunc***$_0$(WORK$_{(N)}$) = [~] proceeds, ***FinFunc***$_0$ (MARKET) = [~] closes.
- ***Caus***(*L*), ***Perm***(*L*) and ***Liqu***(*L*) return verbs that mean respectively ‘to cause’, ‘to let continue’ and ‘to terminate’. They introduce a new actant in the situation, as the subject of the verb: ***CausFunc***$_1$(DISEASE) = give [a ~ to *X*], ***PermFunc***$_0$(ATTACK$_{(N)}$) = condone [an ~], ***LiquOper***$_1$(RESPONSIBILITY) = relieve [*X* from his ~],

Finally, lexical functions can combine in a different way, forming *configurations of functions*. Such functions return collocates that express the meaning of the whole configuration in a single lexical unit and have the syntactic structure controlled by the rightmost component function. For example, ***Magn+Oper***$_1$(DESIRE$_{(N)}$) = burn [with ~].

This concludes my short presentation of lexical functions. In the next section I will discuss a problem particular to the description of Japanese in an ECD. But before that, I provide examples of partial descriptions of collocations controlled by two lexemes of Japanese, SEKININ (‘responsibility/duty’) and USO (‘lie$_{(N)}$’), that illustrate each of the lexical functions introduced[3]. For each collocate, I give the **literal** English meaning in parenthesis; those are **not** proper translations.

[2] I am simplifying a little bit here. See (Mel’čuk 1995) for a full presentation of functions that return light verbs.

[3] All these collocations were found in (Corwin 1968), except ***Magn***(SEKININ). I use the Kunrei system of transliteration.

SEKININ (‘responsibility’)

QSyn	*approximate synonyms*: ninmu (‘task’); sekimu (‘obligation’)
S$_1$	*noun for X*: sekininsya (‘responsibility+person’)
Magn	*big* ~: omoi (‘heavy’), zyudai (‘important), zyuyô (‘important’), ôki (‘big’)
Oper$_1$	*X bears the* ~: [~wo] ou (‘support’)
IncepOper$_1$	*X takes the* ~: [~wo] toru (‘take’)
Func$_1$	*the ~ lies with X*: [*X*ni ~ga] aru
CausFunc$_1$	*to place the ~ on X*: [*X*ni ~wo] tou (‘ask’); [*X*ni ~wo] owaseru (=OU ‘support’ + causative), ositukeru (‘force’), [*X*ni ~wo] tenka suru (‘transfer’)
LiquFunc$_1$	*relieve X of his* ~: [*X*kara ~wo] toku (‘untie’)
Real$_1$	*X carries out his* ~: [~wo] hatasu (‘accomplish’)
AntiReal$_1$	*X does not carry out his* ~: [~wo] kaihi suru (‘avoid’)

USO (‘lie$_{(N)}$’)

S$_1$	*noun for X*: usotuki (‘liar’)
Mult	*pack of* ~: [~no] katamari (‘bunch’)
Magn	*big* ~: makka (‘crimson’)
Ver	*good* ~: mottomo rasii (‘that seems plausible’)
AntiVer	*obvious* ~: miesuita (‘transparent’)
Oper$_1$	*X tells a* ~: [~wo] iu (‘say’), tsuku (‘use’); osieru (‘teach’)
Magn^{quantity}***+Oper***$_1$	*X tells many* ~: [~] happyakuwo naraberu (‘align 800 ~’), [~no] katamariwo naraberu (‘align a bunch of ~’)

4 Japanese numeral classifiers in an ECD

I will now discuss a problem particular to Japanese (as well as other major Asian languages), that of numeral classifiers (or counters, as some would call them) from a lexicographic point of view. Not being a specialist of numeral classifiers nor of Japanese language, I do not pretend to solve their case once and for all; my goal is simply to show how they could possibly be described in an ECD, and I hope to provide good leads for future research. There is already abundant literature on classifiers, so I will introduce them only briefly.

In Japanese, most common nouns need to be preceded by a classifier in order to be quantified with a numeral. For instance, it is not possible to say just *three cats* as such, one has to say something that could be translated literally into *cat in three units*:

(a) san biki+no neko[4]
three CL(assifier)+GEN(itive) cat (*three cats*)

Classifiers such as HIKI are legion (apparently over 200). The choice of the classifier is not free: nouns usually select only one or few, so that one has to remember for each noun what classifier to use with it. Classifiers have often been seen as semantically motivated, and are still sometimes taught as such, because nouns denoting similar objects tend to select the same classifier (for instance, words denoting long, cylindrical objects would usually select HON, those denoting small animals would take HIKI, etc.). However, such classes are not fully consistent. For instance, the same classifier, TYÔ, is used for counting (among other things) tofu, scissors, baskets, candles, rooms, servings of food, city blocks or menus[5]. It would be really far-fetched to pretend that these things form some sort of natural class of objects, and that it is their belonging to that class that dictates the use of TYÔ for their quantifying. In other words, it is not because of their knowledge of the real world that Japanese speakers can select the correct classifier for a given word, but because of their knowledge of the language itself. It is very similar to noun genders in other languages: they can sometimes be deduced from the meaning of words (in Romance languages, for instance, nouns strictly denoting females are usually feminine), but it is only for a limited number of words that such deductions are possible; for all other words, gender is non-motivated and has to be remembered by heart, it cannot be deduced. The same holds for Japanese numeral classifiers. Therefore, a good dictionary should specify, for each noun, which classifier goes with it. Unfortunately, this is far from being the norm in current monolingual and bilingual dictionaries of Japanese.

Then, how to encode the selection of numeral classifiers by nouns in a Japanese dictionary? Where is such information to be put and how can it be formalized? Let’s first have a closer look at the bunch of words we are trying to describe, for it seems that many of them might actually be impostors. If we are to base ourselves on a set of clear criteria, namely their semantic and syntactic behavior and their obligatory nature, we are lead to divide so-called classifiers into two groups: true classifiers, such as HIKI, used to count units or single instances of what is denoted by the noun being quantified (be it a concrete entity or an abstract fact) on the one hand, and

[4] Morphs are separated with a “+” in the examples.

[5] Information found on http://www.trussel.com/jcount.htm.

on the other, nouns denoting measurement units, containers or groups. All of them seem to have a similar syntactic behavior; compare (a) to the following examples:

(b) san rittoru+no mizu
three liter+GEN water (*three liters of water*)

(c) san taru+no mizu
three barrel+GEN water (*three barrels of water)*

(d) huta yama+no hurûtu[6]
two mountain+GEN fruit (*two piles of fruits)*

RITTORU, TARU and YAMA have the same syntax as HIKI. They can all emancipate, so to speak, from their nominal governor and be attached directly to the main verb: *san bikino nekoga iru ~ nekoga san biki iru* ('three CL+GEN cat+SUBJ there-is' ~ 'cat-SUBJ three CL there-is'), *san rittoruno mizuga aru ~ mizuga san rittoru aru* ('three liter+GEN water+SUBJ there-is' ~ 'water-SUBJ three liter there-is'), etc. It is a strange feature enough to let think that these four words belong to the same wordclass. However, there are some non-trivial differences indicating that words denoting measurement units, containers or groups are in point of fact nouns, while words like HIKI are not. First of all, while the former all have a meaning that can be easily glossed, the latter has none: ask any native Japanese speaker what HIKI or TYÔ mean, and you will get a very confused answer. Second, words like RITTORU, TARU and YAMA can all be modified, as any other noun, while those like HIKI cannot:

(e) ippaino san rittoru+no mizu
full+GEN three liter+GEN water (*three full liters of water* — that is, really 3000 ml, not just 2995 ml)

(f) ôkina san taru+no mizu
big three barrel+GEN water (*three big barrels of water)*

(g) ôkina huta yama+no hurûtu
big two mountain+GEN fruit (*two big piles of fruits)*

(h) *ôkina san biki+no neko[7]
big three CL+GEN cat (*three big cats)*

Finally, a noun denoting a measurement unit, a container or a group can be the head of a nominal phrase, while true classifiers cannot. When such a noun is the head of a phrase, it can be quantified, selecting its own true classifier, just like any common noun:

(i) mi+tu+no hurûtu+no yama[8]
three+CL+GEN fruit+GEN mountain (*three piles of fruits)*

(j) *(san+ko/tu/biki/…+no) nekono hiki
(three+CL+GEN) cat+GEN CL (*three cats)*

These semantic and syntactic discrepancies strongly suggest that we are dealing with two different sets of words. This is further reinforced by the obligatory nature of classifiers, as opposed to the free use of nouns. Indeed, it is simply not possible to quantify common nouns without the use of a classifier (except for the very nouns that have been discussed above and perhaps a few exceptions among the common nouns). Classifiers do not bear a meaning that is intended by the speaker; they do not appear in the sentence as a result of a free choice by the speaker. On the contrary, nouns denoting measurement units, containers or groups do have a meaning that is freely chosen by the speaker. They can also be changed at the speaker's will: *san rittoruno / kirono / tonno / ... mizu* ('three liters / kilograms / tons / … of water'); true classifiers cannot. The obligatory nature and the meaninglessness of classifiers such as HIKI suggests that they are grammatical words which are markers of an inflectional category of numerals, though I cannot be sure of it at this point.

Only now that we have taken the impostors out of the bag of classifiers is it possible to answer the question that interests us: how to encode the selection of classifiers in an ECD? But then a second question arises: what to do with the nouns we have found not to be classifiers? Let's answer the second question first, for the sake of suspense.

Words denoting measurement units, containers and groups are all nouns which have the ability to be used as if they were classifiers. I do not know if this ability is common to all of them or if there are exceptions. If it is common to all of them, then it can be deduced from their meaning. Therefore, it is not necessary to put this

[6] For some reason, it seems at least bizarre to most native speakers to say *san/mi yamano hurûtu* ('three mountain+GEN fruit'): YAMA can be used as a classifier only when the numeral is one or two. However, other group nouns could combine freely with any numeral, and the capricious behavior of YAMA actually reinforces the position I am developing.

[7] This phrase is structurally ambiguous. In one of its possible syntactic representations, the adjective ÔKI depends on the noun NEKO; this construction is absolutely correct. However, the construction that interests us is when the adjective modifies not NEKO but HIKI; this interpretation is incorrect. It is not surprising at all, considering that HIKI bears no meaning.

[8] Two remarks should be made about this example. First, TU is a classifier that is selected by the noun YAMA, not HURÛTU. Second, the numeral MI is different from the one used in the other examples (SAN). This is due to the fact that there are two sets of numerals in Japanese for small numbers, and it is the classifier that imposes the use of such or such set. Of course this information has to be written in the dictionary entry of each classifier.

information in their lexicographic entry. If, on the contrary, there are a good number of exceptions and their ability to behave as classifiers or not is not deductible from their meaning, then it is possible to stipulate that there is a part of speech in Japanese, say, "quantity nouns", and that those that have this ability are quantity nouns while those that don't have it are common nouns. This illustrates very well how the study of the lexicon can influence the way we describe the grammar of a language.

Yet, there is one important difference between group nouns and the others: while nouns for measurement units and containers can be freely chosen by the speaker (depending on the intended meaning, of course) it is not the case with nouns denoting groups. The choice of expressing their meaning or not is up to the speaker, unlike true classifiers, but the choice of the lexeme that will express the meaning 'a group of' is dictated or at least restricted by the lexeme that is to be quantified. In other words, they are collocates, and they can be described via a lexical function, namely ***Mult***. For example: ***Mult***(HURŪTU 'fruit') = *[~no] yama* ('mountain'), ***Mult***(HANA 'flower') = *[~(no)] taba* ('bouquet')*, fusa* ('tuft'), ***Mult***(MUSI 'insect') = *[~no] mure* ('group')*, itigun* ('crowd'), etc.

Now, what about the classifiers proper? Expressions like *sanbikino neko* are somewhat reminiscent of phrases such as *piece of information*. HIKI is used to count (units of) cats, therefore couldn't we say that ***Sing***(NEKO) = *hiki [no ~]*, ***Sing***(TÔHU 'tofu') = *tyô [no ~]*, etc.? The use of values of ***Sing*** would simply be mandatory in Japanese for most common nouns, somehow as if most Japanese nouns were considered uncountable. It is a possible solution, indeed. However, there are two important differences between English collocations such as *piece of information* and Japanese classifiers. The first one is that in Japanese, it is the quantified noun that is the syntactic head of the phrase, while usually in typical ***Sing***-type collocations, it is the collocate. Compare the following two phrases (syntactic dependencies are indicated, heads underlined):

(k) san tyô+no □ sentô
three CL+GEN scissors (*three pairs of scissors* — literally, *scissors in three pairs*)

(l) three pairs □ of □ scissors

In addition to this, the obligatory nature of classifiers suggests that they constitute an inflectional category, and that in fact it is a form of agreement. If this analysis turned out to be correct, then it would mean that numerals in Japanese agree in class with the noun they quantify. Therefore, each common noun would have to be assigned a nominal class, exactly as each noun in the ECD of French is assigned a gender.

5 Conclusion

I have briefly introduced the Explanatory-Combinatorial Dictionary, a formalized lexicon which focuses on restricted lexical cooccurrence and constitutes the heart of the Meaning-Text Model of a language. The formal tool used to describe collocations in an ECD, lexical functions, are mathematical functions that return, for a given base, collocates that can be used to express a specific meaning. The examples given show that lexical functions can prove useful for the description of Japanese too. In particular, they offer a very elegant way of describing group nouns, which are traditionally considered as numeral classifiers.

There exists already a multi-lingual lexical database project called *Papillon*, which is based on the *DiCo* project and focuses especially on Japanese and French (Sérasset & Mangeot-Lerebours 2001)[9]. The idea of this project is to develop *DiCo*-like lexica of various languages and then link the lexemes cross-linguistically, yielding a rich, formalized, multi-lingual dictionary. However, the Japanese part of this project is still at an embryonic state.

Acknowledgments

I wish to express my gratitude to Igor Mel'čuk who read and commented draft versions of this article. His remarks and questions were very useful, and he pointed out to me important facts that I might have missed otherwise. Many thanks also to Mami Fukushi and Tomoyo Noguchi, who patiently answered all my silly questions and helped me collecting data. I am of course the only one to blame for all remaining errors and inconsistencies.

[9] The database itself and information about it are available on http://www.papillon-dictionary.org

References

Corwin C (ed.), 1968, *A dictionary of Japanese and English Idiomatic Equivalents*. Tokyo/New York/ London, Kodansha International.

Kahane S & Polguère A, 2001, Formal foundation of lexical functions. In *Workshop proceedings of Collocation: Computational Extraction, Analysis and Exploitation*, 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics, Toulouse, July 7th 2001, pp. 8-15.

Lareau F, 2002, La synthèse automatique de paraphrases comme outil de vérification des dictionnaires et grammaires de type Sens-Texte [Automatic generation of paraphrases as a tool for the verification of Meaning-Text dictionaries and grammars]. M.A. dissertation, Montréal, Université de Montréal.

Mel'čuk I, 1981, Meaning-Text Models: A Recent Trend in Soviet Linguistics. In *Annual Review of Anthropology*, vol. 10, pp. 27-62.

Mel'čuk I, 1995, The Future of the Lexicon in Linguistic Description and the Explanatory-combinatorial Dictionary. In Lee I-H (ed.), *Linguistics in the Morning Calm 3* (Selected Papers from SICOL-1992). Seoul, pp. 181-270.

Mel'čuk I et al, 1984, 1988, 1992, 1999, *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques I, II, III, IV [ECD of French: Lexico-semantic research I, II, III, IV]*. Montréal, Les Presses de l'Université de Montréal.

Milićević J, 1997, Étiquettes sémantiques dans un dictionnaire formalisé du type Dictionnaire Explicatif et Combinatoire [Semantic labels in an ECD-like formalized dictionary]. M.A. dissertation, Montréal, Université de Montréal.

Polguère A, 2000, Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. *Proceedings of EURALEX'2000*, Stuttgart, pp. 517-527.

Polguère A, 2003 (yet to appear at the moment of submitting this article), Étiquetage sémantique des lexies dans la base de données DiCo [Semantic labeling of lexical units in the DiCo database]. In Zock M & Carroll J (eds), *Les dictionnaires électroniques: pour les personnes, les machines ou pour les deux? [Electronic dictionaries: for humans, machines or both?]*. TAL, vol. 44:2.

Sérasset G & Mangeot-Lerebours M, 2001, Papillon Lexical Database Project: Monolingual Dictionaries and Interlingual Links. In *Proceeding of NLPRS'2001*, The 6th Natural Language Processing Pacific Rim Symposium, National Center of Sciences, Tokyo, pp. 119-125.

Wanner L (ed.), 1996, *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam, Benjamins Academic Publishers.

Zholkovsky A & Mel'čuk I, 1984, *Tolkovo-kombinatornyj slovar' russkogo jazyka [Explanatory-combinatorial Dictionary of Modern Russian]*. Vienna, Wiener Slawistischer Almanach.