Formalizing non-standard lexical functions

Anne-Laure Jousse (1)(2)

(1) OLST – Université de Montréal
(2) Lattice- UMR 8094 – Université Paris 7
anne.laure.jousse@umontreal.ca

Abstract

In the lexicographical products developed within the framework of the Explicative and Combinatorial Lexicology, such as the DiCo, the lexico-semantic links are modeled by means of lexical functions. However, only a part of them (called standard) appear as a real formal encoding. The others (called non-standard), which represent links more specific to some lexical units, are written in a heterogeneous and barely formalized way. Therefore, some relations cannot be taken into account in automatic processings. We propose in this paper a methodology for the formalization of non-standard lexical functions in order to make them machine-readable in applications such as text analysis and text generation. To complete this work, we discuss some theoretical assumptions drawn upon this formalism and sketch some propositions for a global and homogeneous formalization of all lexical links described in the DiCo.

Keywords

(non-standard) Lexical Function, modeling of lexico-semantic links, formalization, DiCo

1 Introduction

In various NLP applications, it is necessary to have a formal system to describe the idiomatic lexical-semantic relations between the lexemes of a language. By lexical-semantic relations, we mean syntagmatic relations or collocations (e.g.: for bee: to sting, to gather nectar, to pollinate...) as well as relations of paradigmatic nature, identified in the Explicative and Combinatorial Lexicology as semantic derivatives (e.g.: for bee: hive, swarm, beekeeper...). In text understanding, one objective of the modeling of these lexical relations is to allow inferences starting from a lexeme: for example, with an encoding of the relation between an animal and its home, one can infer, starting from the lexical unit hive, the lexical unit bee, starting from stable, horse, etc. On the contrary, in text generation, a good description of lexical links makes it possible to ensure relative idiocy of generated texts. For example, to express the relation linking a lexeme indicating an object and the mark left by this object, one will be able to use an appropriate noun: for a ship, one will speak about wake; for a tire or a
step, of trace or print; for a wound, of scar or gash, etc. The Meaning-Text Theory provides, by the means of lexical functions, a good modeling of lexical-semantic relations. They encode the meaning and the syntactic characteristics of a relation in short formulas taking a key word and returning one or more elements of the value (according to the following pattern: **key word LF value**). For example:

- Typical name of place: bee, hive
- Typical verb of realization: chisel, to carve
- Positive qualifier: desire, noble

Lexical functions (from now on LF) have already been largely used in NLP applications (Apresjan et al. 2000), for paraphrasing and text generation (Nasr 1996, Lareau 2002), automatic abstracting (Kittredge & Bélanger 2005), machine translation (Apresjan 2003), as well as for the development of language learning tools (Diachenko 2006, Boguslavsky et al. 2006). However, not all LFs are used in these projects. Indeed, the encoding of LFs is not homogeneous but varies according to the degree of frequency of lexical relations. The larger the set of lexemes to which an LF is applicable, the more it will be regarded as standard. Thus, whereas the relation modeled by $S_{loc}$ can be applied to many lexemes, the relation between criminal (=criminal) and en cavale (=on the run) which can be described by ‘Qui cherche à échapper à la police’ (=’who is fleeing from the police’) only relates to a very small number of lexemes (criminal, gangster, murderer, etc). These relations, specific to certain lexemes, will be described by means of non-standard LFs (or NSLFs). They are encoded according to a natural metalanguage which emphasizes legibility and clearness of formulation. This has as a consequence the existence of synonymic alternatives, as illustrated in the three following examples:

| médicament | pharmacologie | Pharmacology which studies M. |
| estomac | gastroentérolgie | Gastroenterology which deals with affections of ~ |
| poisson | ichthyologie | Ichthyology which studies F. |

As such, this encoding is very heterogeneous and not exploitable for automatic processing since it can not be spotted or analyzed automatically. Thus it is necessary to formalize this encoding and there lies the objective of our work. Our goal is to create a formal encoding that can be automatically processed and improves the organization and the way of accessing lexical relations in a DiCo-like dictionary. We illustrate in this article our method using the nominal and adjectival relations extracted from the DiCo (French lexical database developed at the OLST, by A. Polguère and I. Melčuk). The first part of the article is devoted to a theoretical discussion of the status and legitimacy of a distinction between lexical-semantic relations. The second part accounts for the work of formalization of nominal and adjectival non-standard LFs.

---

1 In order to lighten the text we don’t exactly follow the DiCo’s formalism.
2 Theoretical issues for a formalization of non-standard lexical functions (NSLFs)

2.1 Definition and problems

The aim of LFs is to describe a lexical-semantic relation by a synthetic encoding, accounting for the meaning and the syntactic characteristics of the relation. For example, the LF Real, applies to nouns and returns a verb with a meaning of realization; it takes as its object the key word of the relation, e.g. car Real, to drive [a car]. However, as we have briefly mentioned, the LFs are not homogeneous. They are organized around three levels of “standardness” that start from a very synthetic and general description (standard LFs) and go to a very precise description (NSLFs). Let us note on this subject that these statuses are not hermetic but are rather located on a continuum.

a) The standard LFs form a closed set of about 60 (Mel'čuk et al., 1995:125), they must comply with the principle of universality, i.e., they can be found in every natural language. They are made up of only one element (see 1) or may be combined with some of them (see 2).

(1) happy Anti sad : 'antonymy'
(2) abscess IncepPredPlus to mature : 'to become to be more intense'

b) A semi-standard LFs consists of a standard LF and an element written in a natural language that adds a component of meaning not dealt with by the standard LF. For example, the description of the relation (3), which means 'to cause the silence of someone', differs from the relation (4) by the presence of the element en échange de qqch (= in exchange for sth), which brings an additional component to the meaning of the standard LF.

(3) silence CausOper1 réduire [qqn. au silence] (= to make s.o. keep silent)
(4) silence en échange de qqch. CausOper1 acheter [le silence de qqn.](=to buy s.o.’s silence)

c) The non-standard LFs (NSLFs) are completely written in the language of description.

(5) canard Qui ne vit pas à l’état sauvage d’élevage, domestique
(=duck Which does not live in a wild state domestic)
(6) bière Produite à l’étranger d’importation
(=beer Produced abroad imported)

Polgùère (2007: 4) suggests to assign the standard status to the LFs satisfying the two following conditions:

- cardinality condition: The LFs must apply to a significant number of lexemes and not be limited to the lexemes of only one semantic field. For example, the relation (5) can only apply to a few names of animals and thus will not be able to claim a standard status.

- diversity condition: The values returned by the LFs must be diversified. For example, the values of the LFs Bon ‘positive qualifier’ are diversified according to the keywords it takes: aspiration Bon noble; bijou (=jewel) Bon somptueux (=sumptuous); déjeuner (=lunch) Bon succulent (=scrumptious), etc. Consequently, one can regard it as a standard LF.
These conditions offer important criteria for the description of the relations but do not always appear sufficient. Without completely putting this distinction into question, we would like to consider its legitimacy and its relevance with respect to the inherent aims of the LFs. The fundamental problem that we raise here is as follows: it seems that the way of encoding the non-standard relations contributes to relegate a considerable set of relations to the margins of the formal descriptions carried out in the DiCo. The NSLFs account for 16% of the relations of the DiCo (1 931 non-standard out of 11 912 relations), but if one wants to model the lexicon by means of a formal encoding, why would one leave aside a relatively great part of the lexico-semantic relations? We wish to call into question, on the one hand, the allegedly fixed number of standard relations, and on the other hand, the way in which the NSLFs are encoded. Whereas the number of simple standard LFs is fixed at about sixty, the NSLFs constitute an open set of semantic relations. Determining a number of LFs is necessary to preserve coherence and a reasonable maintenance. However, a too peremptory limitation could stand in the way of a good development of the LFs’ system. It seems to us that if the creation of new standard LFs permits the removal of a great number of NSLFs, it is perhaps legitimate to add some of them to the list. In addition, the strict distinction between standard and non-standard LFs has involved a radically different way to treat the lexical-semantic relations. We think that limiting the number of the standard LFs shall not prevent a rigorous formalization of the other relations. Unfortunately, in the current state of the DiCo, there is no known attempt to formalize NSLFs. Some regularities, although obvious, are not taken into account. In order to make up for this problem, Polguère (2007: 4) proposed another category of LFs with a hybrid status (between standard and non-standard) for the NSLFs whose universal nature has not been yet demonstrated but which can be called standard for a given natural language. They are called local standard functions and are written in formulas of the described language rather than in Latin. Thus, although the non-standard function De_nouveau (=Again) illustrated below,

\[
\text{hostilité (=} \text{hostility}) \quad \text{De_nouveauIncepFunc}_0 \quad \text{reprendre (=} \text{to begin again})
\]

\[
\text{goût (=} \text{taste}) \quad \text{De_nouveauIncepOper}_{12} \quad \text{reprendre \{goût à qqch.\} (=} \text{to do sth. again})
\]

is very frequent in French, it is non-standard because its universality remains to be proven. However, it is regarded as a local standard lexical function of French. In our view, this initiative seems to be a good solution to reduce the number of NSLFs. Unfortunately, this principle has not truly been applied to other LFs than those described in the quoted article.

### 2.2 Towards a formal and balanced level of description: which scale of granularity shall be chosen?

The central question when one seeks to describe the meaning of a relation is to choose an ideal degree of accuracy. We would like to introduce here the concept of granularity in the description of a relation between lexemes. As Polguère stated under the conditions referred to above, the granularity of the relation shall not be too restricted, i.e. the description of the relations must be sufficiently broad to relate to a large set of lexical items. However, no criterion is defined for the contrary case, namely, to avoid a too general description of a relation. In this view, the challenge is to find a good balance between a too fine and too broad degree of granularity. If one chooses a very precise description, the number of formulas will grow. On the other hand, if one aims for a maximal generalization, the clearness of the relation will be jeopardized. Let us consider the examples of Figure 1 below. All of them
illustrate a relation between an entity and a product it is derived from. We could choose to gather the names of animals and the courses prepared with their flesh. In this case, the selected LF would reflect a fine granularity which would exclude the relations between elephant and ivory, or bread and slice of bread, etc. On the contrary, pitching on a broader description like Produit dérivé (=derived product) makes it possible to gather all of them under a unique formula.

<table>
<thead>
<tr>
<th>Produit dérivé</th>
<th>Foie de C. produit par gavage</th>
<th>foie gras</th>
</tr>
</thead>
<tbody>
<tr>
<td>(= duck) sucre</td>
<td>Liver produced by force-feeding of D.</td>
<td>foie gras</td>
</tr>
<tr>
<td>(= sugar) cerise</td>
<td>Friandise contenant beaucoup de S.</td>
<td>sucrerie</td>
</tr>
<tr>
<td>(= cherry) éléphant</td>
<td>Delicacy with lots of S.</td>
<td>confectionary</td>
</tr>
<tr>
<td>(= elephant) lait</td>
<td>Aliment préparé dérivé du L.</td>
<td>cherry, kirsch</td>
</tr>
<tr>
<td>(= milk) pain</td>
<td>Matériau que l’on obtient à partir des défenses d’É.</td>
<td>ivory</td>
</tr>
<tr>
<td>(= bread) pain</td>
<td>Boisson alcoolisée faite avec des C.</td>
<td>cherry brandy</td>
</tr>
<tr>
<td></td>
<td>Alcoholic beverage made of C.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sucre</td>
<td>Friandise contenant beaucoup de S.</td>
</tr>
<tr>
<td></td>
<td>cerise</td>
<td>Aliment préparé dérivé du L.</td>
</tr>
<tr>
<td></td>
<td>éléphant</td>
<td>Matériau que l’on obtient à partir des défenses d’É.</td>
</tr>
<tr>
<td></td>
<td>lait</td>
<td>Aliment préparé dérivé du L.</td>
</tr>
<tr>
<td></td>
<td>pain</td>
<td>Tranche de P. enduite de N</td>
</tr>
<tr>
<td></td>
<td>pain</td>
<td>Peace of B. spread with N</td>
</tr>
</tbody>
</table>

Figure 1: Homogenization of the encoding of nominal NSLFs

In the DiCo, each standard LF has a popularization formula that makes the LF explicit. For example, the following relation: gare (=station) IncepReal1 arriver, entrer [en ~] (=come [into ART ~]) is popularized by ‘[X] arriver dans une G.’ (‘[X] to arrive into a S.’). We wish to have, for the NSLF, a maximal generalisation of the relation with a systematic verbal formula just as the standard LF. Thus, one will have both a formal encoding for automatic processings and an explicit formula for human users.

3 Proposals for the formalization of NSLFs

A certain amount of research has been undertaken to propose new lexical functions and standardize the NSLFs. Erastov (1969, quoted in Polgure 2007), Grimes (1990) and Fontenelle (1997) suggested some new lexical functions. Frawley (1998), L’Homme (2002), Jousse (2002), Jousse & Bouveret (2003) and Bouveret (2006) postulated some adaptations of the lexical functions to terminological relations. Grizolle (2003) and Jousse (2003) thought of means for the homogenization of some NSLFs. Popovic (2003) worked on the homogenization of the standard LFs’ verbal formulations. However, concerning the question of the encoding of the NSLFs as a whole, no concrete solution has been proposed yet. We wish to propose a guideline for a rigorous formalization of the NSLFs. One must draw a distinction between standardizing and formalizing NSLFs. While the former aims at creating new standard LFs, which involves confronting candidates to the previously mentioned criteria, the latter aims at observing regularities among lexical relations and proposing a formalization of their encoding without thinking of them as standard LFs. Thus, formalization can be regarded as the first step toward standardization, when this is possible.

As a method, we will distinguish the relations according to their parts of speech, that is to say: nominal, adjectival, verbal and adverbial relations. Let us note that, in this article, we will deal only with the first two types of relations. Our corpus can be measured in terms of LFs:
we have extracted the 820 nominal NSLFs as well as the 727 adjectival NSLFs from the DiCo.

### 3.1 Standardization and formalization of nominal NSLFs

The relations encoded by nominal NSLFs are mainly semantic derivatives. They are lexemes sharing a component of meaning with a key word (thus, *beekeeper* is a semantic derivative of *bee* because it is defined as a 'person raising bees'). We have looked for regularities among the nominal NSLFs while taking into account the former proposals made by other linguists. It is our belief that amongst the LFs proposed below, some could claim to be standard, as for the others, they will be used as models for the formalization of the NSLFs.

#### 3.1.1 The *Matériau* and *Ingrédient* relations

The 'typical material' and 'typical ingredient' relations are sometimes regarded as meronymy. For example, one finds in WordNet (Fellbaum 1998) three types of meronymy: member meronymy (association → associate), substance meronymy (steel → iron) and part meronymy (table → leg). In order to encode these relations, we propose the two following functions: *Matériau* and *Ingrédient*, which could be standardized as well as the traditional meronymy because they fulfill the necessary conditions to this status.

<table>
<thead>
<tr>
<th>Matériau</th>
<th>Ingrédient</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>assiette</strong></td>
<td>Matériau typique dont sont faites les A.</td>
</tr>
<tr>
<td><strong>corde</strong></td>
<td>Matériau pour C.</td>
</tr>
<tr>
<td>(= Material)</td>
<td>porcelaine</td>
</tr>
<tr>
<td><strong>plate</strong></td>
<td>Typical material of which P. are made</td>
</tr>
<tr>
<td><strong>rope</strong></td>
<td>Material for R.</td>
</tr>
<tr>
<td></td>
<td>lin, coton, nylon</td>
</tr>
<tr>
<td><strong>bière</strong></td>
<td>Ingrédient utilisé pour faire de la B.</td>
</tr>
<tr>
<td><strong>yaourt</strong></td>
<td>Ingrédient avec lequel on fait le Y.</td>
</tr>
<tr>
<td>(= Ingredient)</td>
<td>blé, houblon, malt</td>
</tr>
<tr>
<td><strong>beer</strong></td>
<td>Ingredient used to make B.</td>
</tr>
<tr>
<td><strong>yogurt</strong></td>
<td>Ingredient of which one makes B.</td>
</tr>
<tr>
<td></td>
<td>corn, hop, malt</td>
</tr>
<tr>
<td></td>
<td>fermented milk</td>
</tr>
</tbody>
</table>

Figure 2: *Matériau* and *Ingrédient*

#### 3.1.2 The relations Masc, Fem, Infant

It is difficult to claim that the three following relations comply with standardness insofar as they cannot apply themselves other than to a restricted number of lexemes: names of individuals, trade and animals. However, their encoding being unnecessarily heterogeneous, we propose to gather them under three single formulas: *Masc* for the male equivalents, *Fem* for the female equivalents and *Infant* for the names of young animals or individuals (see Figure 3).
Formalizing non-standard lexical functions

3.1.3 The relation between something and the scientific discipline which studies it: Schol

In many cases, the relation between a fact or an entity and its scientific study is morphologically marked by the element -logie in French or -logy in English, which demonstrates a visible recurrence of the relation (see examples of this relation given in the introduction). Studies on the relation between morphological derivatives and standard LFs (Jousse 2002, Jousse and Bouveret 2003) showed that the majority of typical morphological derivatives of French are already modeled by LFs. For the treatment of certain NSLFs, it seems to us relevant to look at the most current greco-latin compositions as a reference for the regularity of certain lexical relations.

For lack of space, we cannot make a detailed inventory of all the NSLFs which we formalized. In the current state of our work, we succeeded in treating 720 nominal NSLFs (out of the initial 820) which we divided into 30 formalized NSLFs. We must admit that the remaining relations are not easily generalizable. For example, the relation between chat (=cat) and chatière (=cat flap), represented by the NSLFs Petite ouverture pratiquée en bas d'une porte qui permet à un C. d'entrer et de sortir (=Small opening made in the bottom of a door which makes it possible for C. to enter and exit). However we specify that the DiCo is still very little developed (1000 separated senses of lexical items) which makes it more difficult to extract regularities. As this work is correlated with the advancement of the DiCo, it cannot be made a priori and it is obvious that the data will evolve progressively with the description of new lexical relations.

3.2 Formalization of adjectival NSLFs

To treat the adjectival NSLFs we adopt a different approach insofar as the relations concerned are not of the same order. The relations encoded by the adjectival NSLFs are mostly of the syntagmatic type (of collocations) taking names for key words and returning adjectival values. For example: victoire Dont la probabilité est faible douteuse (=victory Whose probability is weak doubtful). In order to ensure a coherent and global treatment of the adjectival NSLFs, we make the assumption that any type of entity or fact has got attributes, for example: 'function', 'size', 'shape', 'appreciation', etc., that are likely to be expressed by means of adjectival collocatives. These attributes can be compared with the Frame Elements described in FrameNet (Fillmore et al. 2003), which specifies the elements entering a conceptual frame. For example, under the Artifact frame, one finds the elements ‘creator’,

<table>
<thead>
<tr>
<th>Masc</th>
<th>Fem</th>
<th>Infant</th>
</tr>
</thead>
<tbody>
<tr>
<td>abeille</td>
<td>Mâle</td>
<td>Petit du C.</td>
</tr>
<tr>
<td>mouton</td>
<td>M. mâle</td>
<td>chaton</td>
</tr>
<tr>
<td></td>
<td>faux bourdon</td>
<td>(=bee Male drone)</td>
</tr>
<tr>
<td></td>
<td>bélier</td>
<td>(= sheep Male S. ram)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avocat</td>
<td>De sexe féminin</td>
<td>avocat</td>
</tr>
<tr>
<td>chien</td>
<td>Femelle</td>
<td>chienne</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(=lawyer Female lawyer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(=dog Female bitch)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chat</td>
<td>Coq du C.</td>
<td>coq</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(=cat Baby C. cockerel)</td>
</tr>
</tbody>
</table>
Anne-Laure Jousse

‘material’, ‘name’, ‘time of creation’, ‘type’, ‘use’,... These elements of conceptual nature are then likely to be lexicalized in sentences dealing with a concept of artifact. According to the analysis of the various NSLFs, we found a great number of attributes. We then assigned an attribute to each NSLFs in order to propose a first homogeneous formalization of it. This new type of encoding makes it possible to organize and to reach them more easily in the DiCo. Figure 6 presents some of these attributes according to the following pattern: old LFs → standardized LFs)

**COULEUR (= COLOR)**
- *barbe* De couleur grise → Couleur: gris
- *poivre et sel* (beard: gray pepper-and-salt)

**FORME (= SHAPE)**
- *barbe* Qui a une forme évasée → Forme: évasée en éventail
- *poivre et sel* (beard: splay)

**TAILLE (= SIZE)**
- *drap* Utilisable pour deux personnes → Taille : + double
- *poivre et sel* (sheet: double)

**MATÉRIAU (= MATERIAL)**
- *chapeau* Qui est fait de paille → Matériau: paille de paille
- *poivre et sel* (hat: straw)

**FONCTION (= FUNCTION)**
- *wagon* Équipé pour que les passagers y dorment → Fonction: dormir ~lit
- *poivre et sel* (railcar: sleeper)

**CAUSE (= CAUSE)**
- *célibat* Qui a lieu malgré la volonté de X → Cause : non voulu forcé, obligé
- *poivre et sel* (celibacy: unwanted compelled)

**CONSEQUENCE (= CONSEQUENCE)**
- *coup de feu* Qui tue l’être Y → Conséquence: mort de Y mortel
- *poivre et sel* (gunshot: kill Y mortal)

Figure 6: Some examples of normalized adjectival NSLFs

Some of these attributes are polarizable, for example: Size:+ or Size:--; others require more elaborated descriptions to be understood, for example: Functioning: with a chime or Function: to sleep. The list presented above is not exhaustive, it simply aims to illustrate our approach. We have successfully classified all adjectival NSLFs with about 20 attributes while trying to restrict this number as much as possible. However, we are aware that certain groupings were forced a little. Following the example of nominal NSLFs, we recall that this work cannot be claimed accomplished since it will be subject to a constant evolution with the development of the DiCo.

4 Conclusion

In analyzing the regularities among the semantic lexical relations described in the DiCo, we proposed to formalize the non-standard lexical functions. We called into question certain principles related to the lexical functions with the aim of considering new possibilities for their encoding. A formalization is necessary if one wants to obtain a homogeneous formal representation of all the lexical relations and use them in automatic processing. Once formalized, NSLFs will be able to enter a semantic typology of lexical relations which we are
developing. This typology will allow searching for lexemes by their meaning when using the DiCo. The attributes will be of great importance here because of their potential of categorization. For example, one will be able to obtain all the lexemes that deal with the aspect of an artefact, or with its way of functioning, etc. This work constitutes a first stage given that we have only taken into account the nominal and adjectival relations. We must also specify that this is an evolutionary task parallel with the development of the DiCo and that the results will be brought to certain changes. It thus seemed to us paramount to pose a precise methodological framework for the continuation of our objectives.

Acknowledgements

I would like to thank Frédéric Osterrath, André Bittar, Tilmann Reuther and the three anonymous reviewers for their comments which helped me to improve this paper.

Bibliography


Lareau F. 2002. *La synthèse automatique de paraphrases comme outil de vérification des dictionnaires et grammaires de type Sens-Texte*. Mémoire de maîtrise, Université de Montréal.


