# La BDéf : base de définitions dérivée du *Dictionnaire explicatif et combinatoire*

Joel Altman et Alain Polguère

OLST — Département de linguistique et de traduction
Université de Montréal, C.P. 6128, succ. Centre-ville
Montréal (Québec) H3C 3J7 CANADA
joelaltman@hotmail.com alain.polguere@umontreal.ca

## Résumé — Abstract

La base de données BDéf, en cours de construction, contient les définitions lexicographiques des quatre volumes publiés du *Dictionnaire explicatif et combinatoire du français contemporain* (DEC), encodées dans un formalisme rigoureusement défini. La finalité de la BDéf est double. Il s'agit, d'une part, de rendre disponible pour la recherche en sémantique computationnelle un ensemble représentatif de définitions lexicales Sens-Texte. D'autre part, le travail de construction de la BDéf permet de mener une recherche sur la structure interne des sens lexicaux et sur la façon dont celle-ci doit être modélisée. Il s'agit donc d'élaborer un métalangage formel de définition. Cet article comprend quatre sections : problématique générale de notre travail, caractérisation de la BDéf par opposition aux définitions des DEC publiés, méthode d'encodage utilisée puis état d'avancement de la BDéf et débouchés escomptés.

We introduce the BDéf, a formal database of lexicographic definitions fully derived from the four volumes of the *Explanatory Combinatorial Dictionary of Contemporary French* (ECD). The BDéf has two aims. Firstly, it will make a representative subset of formal lexicographic definitions available for research in computational semantics. Secondly, building the BDéf allows us to conduct a much needed research on the internal structuring of lexical meanings and on how it should be modelized. In other words, by constructing the BDéf, we arrive at a formal metalanguage for lexicographic definitions. This paper is divided into four sections : general presentation of the research, characterization of the BDéf as opposed to standard ECD definitions, encoding methodology and, finally, current status and expected use of the BDéf.

## Mots-clés — Keywords

Lexicologie explicative et combinatoire, définition lexicographique, lexicographie formelle.
Explanatory and combinatorial lexicology, lexicographic definition, formal lexicography.

## 1 Introduction

Jusqu'à ce jour, quatre volumes du *Dictionnaire explicatif et combinatoire* (DEC) ont été publiés en format papier — voir (Mel'čuk *et al.*, 1984, 1988, 1992, 1999). Pris ensemble, ces quatre volumes contiennent les articles de 510 vocables français ; ils nous fournissent 1 583

définitions lexicales. (On a donc dans le DEC une moyenne approximative de trois lexies par vocable.) Il s'agit bien entendu d'une modélisation très partielle du lexique français ; cependant, la grande finesse des descriptions sémantiques de la lexicologie explicative et combinatoire (LEC) fait de l'ensemble des informations contenues dans les DEC une base de connaissances sur la sémantique lexicale de grande valeur pour la recherche. On pourrait alors légitimement se poser la question de savoir pourquoi les DEC publiés ne sont que très marginalement utilisés dans le cadre de travaux de recherche sémantique, même au sein de l'Observatoire de linguistique Sens-Texte (OLST). L'explication est à notre avis évidente et est liée aux deux problèmes suivants posés par les DEC publiés.

Premièrement, la théorie de la LEC et la technique lexicographique ont évoluées depuis les premiers travaux effectués lors de la publication du DEC I. La LEC n'a cessé de s'affiner, ce qui fait que nous avons, du DEC I au DEC IV, un ensemble de descriptions lexicographiques relativement hétérogène. Bien entendu, on reconnaît d'un volume à l'autre la même approche, la même structuration des articles et, dans l'ensemble, les mêmes formalismes. Cependant, beaucoup de changements ponctuels ont été apportés dans la façon de construire et d'encoder les définitions, modifications qui justifieraient une mise à jour de toutes les définitions antérieures au DEC IV, et même des définitions de ce dernier volume. Deuxièmement, aussi étrange qu'il puisse paraître pour un projet lexicographique effectué dans le cadre d'une approche théorique ayant ses racines dans le traitement automatique de la langue, il n'existe aucune version informatisée fiable du DEC ! On ne dispose que des fichiers Word originaux, fichiers ne contenant même pas les nombreuses corrections effectuées sur les épreuves finales.

En résumé : le DEC n'existe pas sous une forme informatisée uniforme, directement accessible pour la recherche. On peut envisager deux façons de résoudre ce problème : soit l'on tente de récupérer automatiquement les anciennes données du DEC pour les stocker dans un nouveau format, soit on reprend tout à la main en faisant un grand « ménage de printemps ». Des expérimentations sur la première solution ont été faites dans le passé (Sérasset, Polguère, 1997). Elles n'ont pas permis de récupérer toutes les données et, surtout, elles n'ont pas donné lieu à la création d'un outil véritablement opérationnel de construction et de gestion des données lexicales du DEC, outil qui permettrait notamment d'effectuer le travail indispensable de révision. Cet échec relatif est dû en grande partie au manque d'homogénéité dans l'encodage. Il est aussi dû au fait que les fichiers disponibles n'ont pas été formatés de façon rigoureuse et contiennent une multitude d'idiosyncrasies de formatage qui en rendent la récupération automatique très fastidieuse. La seconde solution s'impose donc : il faut reprendre manuellement les données du DEC et les encoder dans un autre format, entièrement manipulable au moyen de logiciels simples d'analyse de chaînes de caractères.

Si on la considère uniquement sous cet angle, une telle tâche présente en elle-même peu d'intérêt de recherche, si ce n'est de fournir une ressource très riche sur la sémantique lexicale du français. Cependant, le travail que nous présentons ici va bien au-delà de la seule révision et homogénéisation des définitions existantes du DEC. Il s'agit aussi et surtout d'effectuer une étude systématique de la façon dont doivent être structurées et encodées formellement les définitions du DEC pour rendre compte de l'organisation interne des sens lexicaux. Une telle modélisation formelle permettra de traiter automatiquement les définitions et d'effectuer de véritables « calculs sémantiques», par exemple :

1. À partir de la définition de deux lexies, on doit être capable de calculer automatiquement leur proximité sémantique. Que ces deux lexies soient ou non des quasi-synonymes, on doit pouvoir déterminer ce qui les distingue et donc déterminer quelles

conditions doivent être remplies dans un contexte d'énonciation donné pour effectuer telle ou telle lexicalisation — sur une stratégie de lexicalisation fondée sur les lexiques Sens-Texte, voir (Polguère, 1998) ou, en anglais, (Polguère, 2000a).

2. Les définitions ainsi formalisées devraient nous permettre d'extraire de façon automatique ou semi automatique des corrélations entre des caractéristiques sémantiques des lexies et certaines de leurs propriétés de combinatoire ; nous faisons ici référence aux connexions qui peuvent exister entre la définition et les liens de fonctions lexicales syntagmatiques contrôlés par la lexie vedette d'un article de DEC.

3. Une base de définitions formalisées devrait aussi permettre de faire des expérimentations sur le paraphrasage, tel qu'envisagé par exemple dans (Milićević, 2003).

C'est dans ce contexte que nous avons entrepris de développer une base de données de définitions lexicales dérivée des définitions du DEC publié : la BDéf. Le présent article décrit la BDéf dans son état actuel. Nous présenterons tout d'abord une caractérisation générale de la BDéf (section 2) ; ensuite, nous expliciterons les notions descriptives de base que nous avons utilisées pour structurer cette base (section 3) ; finalement, nous conclurons en présentant l'état d'avancement de nos travaux ainsi que les développements à venir (section 4).

# 2 Propriétés générales du modèle descriptif de la BDéf

## 2.1 Rappel de la structure générale des définitions du DEC

Avant d'entamer la description de la BDéf, il est utile de rappeler brièvement comment se présente une définition de DEC standard. Elle se divise en deux parties principales :

1. le *definiendum*, qui est une proposition élémentaire linéaire présentant la lexie vedette accompagnée de ses actants sémantiques ;

2. le *definiens*, qui est la définition proprement dite, c'est-à-dire une paraphrase du definiendum structurée autour d'un noyau sémantique classifiant accompagné de composantes sémantiques distinctives.

Une définition du DEC appartient donc à la famille « traditionnelle » des *définitions par genre prochain et différences spécifiques*. Une définition de ce type présente une décomposition du sens lexical en termes de sens plus simples le constituant, c'est-à-dire une analyse du sens lexical. Pour cette raison, (Polguère, 2003a) propose de la désigner par le terme de *définition analytique*. Voici, à fin d'illustration, la définition de la lexie DÉFIER**I.1** [*Marcel a défié ce pédant en duel à l'épée.*] du DEC IV :

(1) *X défie Y en W à Z* = X croyant que Y a insulté X ou ⌜a porté atteinte⌝ à l'honneur de X, || X communique à Y que X veut que Y prenne part dans un combat W contre X en utilisant l'arme Z, dans le but de punir Y — pour ce que Y a fait à X — en blessant**I.1** Y ou en tuant Y

Comme les définitions du DEC son bien connues dans leur formalisation aussi bien qu'au niveau des principes théoriques qui les sous-tendent, nous présupposerons qu'il est inutile de les introduire à nouveau dans le présent article et nous nous engagerons sans plus attendre dans la description de la BDéf.

## 2.2 Écart formel entre les définitions de la BDéf et celles des DEC publiés

Rappelons que le but ultime de notre recherche est de construire une base de données de définitions de type DEC répondant à une stricte modélisation formelle. Cette base de données se limite dans un premier temps à la nomenclature des quatre volumes déjà publiés du DEC français. Elle doit être utilisable dans le cadre des trois tâches suivantes :

1. permettre la consultation des définitions par des utilisateurs « humains » et servir de support pour la rédaction de nouvelles définitions par les lexicographes du DEC ;

2. fournir des données sur la sémantique lexicale qui soient exploitables par les lexicologues et sémanticiens dans le cadre de recherches de nature théorique ;

3. être compilable en un ensemble de tables pouvant être consultées par des programmes informatiques, dans le cadre d'applications en traitement automatique de la langue.

Pour satisfaire les besoins variés de tous ces utilisateurs potentiels, le formalisme d'encodage des représentations stockées dans la base doit répondre à trois critères.

Premièrement, il doit être interprétable autant par un humain que par un ordinateur. Bien qu'utilisant la langue naturelle (lisibilité par l'humain), il nous faut éliminer toute synonymie et expliciter tout ce qui est exprimé de façon cumulative et idiomatique dans les définitions linéaires du DEC. Notre modélisation vise une correspondance unique entre chaque unité sémantique et son expression dans notre représentation. C'est dans cette visée que nous développons un lexique et une grammaire pour notre métalangage sémantique. Deuxièmement, la structuration des composantes de la définition doit faire partie intégrante de la représentation même. Il nous faut expliciter l'anatomie sémantique de chaque lexie vedette non seulement pour en comprendre le sens dans son ensemble, mais aussi pour comparer ce dernier aux autres sens lexicaux de la langue et pour rendre compte d'une grande variété de phénomènes tels que l'héritage lexical, la lexicalisation en synthèse de texte, la relation entre la définition de la lexie vedette et les liens de fonctions lexicales qu'elle contrôle, etc. Troisièmement, la représentation ne doit pas être soumise aux contraintes qu'impose l'encodage linéaire. Elle doit avoir la puissance formelle d'un encodage multidimensionnel du type réseau sémantique connexe. Ce critère semble aller à l'encontre des deux premiers : comment peut-on avoir une représentation non linéaire qui soit aussi facilement interprétable par l'humain que l'est un encodage fondé sur un métalangage linéaire ? C'est là toute l'originalité de notre approche. En combinant des éléments du langage formel des réseaux sémantiques et des propositions élémentaires logiques avec un métalangage sémantique sophistiqué, nous pensons être parvenus (ou être en voie de parvenir) à une représentation très souple, manipulable par l'humain comme par la machine et qui rende explicite la structure des sens lexicaux.

En un sens, la formalisation que nous utilisons joue, au niveau de la définition, le même rôle que la formalisation utilisée pour les autres éléments de la description lexicale dans le cadre du DiCo (Polguère, 2000b, 2000c) : la structure d'une entrée de DiCo se présente comme un « texte lexicographique ». En tant que texte, il est particulièrement adapté pour supporter le travail de rédaction lexicographique de même que le passage vers des encodages plus destinés au grand public comme le *Lexique actif du français* (LAF) — voir références ci-dessus. En tant que texte entièrement formalisé, l'entrée de DiCo peut être compilée automatiquement en tables de données directement exploitables par un programme informatique. La base de données de définitions que nous construisons possède toutes ces propriétés. Notre approche de

l'encodage des définitions repose entièrement sur les notions de base de la LEC, telles que présentées dans (Mel'čuk *et al.*, 1995). Nous allons prendre ces notions pour acquises et nous concentrer, dans ce qui suit, sur la présentation de ce qui fait l'originalité de l'encodage propositionnel.

# 3 Notions descriptives de base

Les notions de base présentées ici servent à modéliser la structure interne des définitions formalisées de la BDéf. Elles réfèrent à des éléments de la structure de ces définitions, éléments plus ou moins proches de l'unité de base de la définition, que nous appelons *proposition élémentaire définitionnelle*. Tous les éléments décrits ici sont qualifiés par l'adjectif *définitionnel*, qui signifie donc dans notre terminologie ‹qui a rapport à la structure d'une définition formalisée›. Passons maintenant à la présentation des notions centrales suivantes : proposition élémentaire définitionnelle (section 3.1) et bloc définitionnel (section 3.2).

## 3.1 Proposition élémentaire définitionnelle

La BDéf est constituée d'un ensemble de définitions analytiques construites selon les principes de la LEC et encodées sous forme « atomisée » : chaque définition se présente sous la forme d'un ensemble structuré de propositions élémentaires. La proposition élémentaire est l'élément syntaxique de base de notre métalangage descriptif. À quelques rares exceptions près, tout élément plus petit que la proposition élémentaire relèvera du lexique de ce métalangage. Notons que l'approche qui nous avons adoptée est fortement inspirée — pour ce qui est de l'explicitation des propositions élémentaires et de la structuration des définitions en blocs de sens clairement identifiés — par les structures définitionnelles utilisées par A. Wierzbicka (Wierzbicka, 1985, 1987). Nous ne remettons cependant pas en question le principe du « bloc maximal », qui veut que les définitions offrent une décomposition minimale du sens, et nous ne cherchons pas à structurer nos propositions élémentaires autour de l'usage de primitifs sémantiques, comme le fait Wierzbicka.

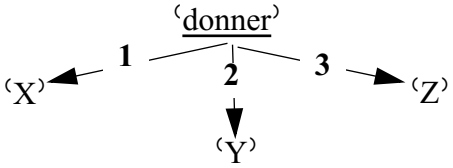
### *3.1.1 Caractérisation formelle de la proposition élémentaire*

La proposition élémentaire se caractérise par son contenu et par sa forme. Du point de vue de son contenu, la proposition élémentaire équivaut à une proposition logique. C'est-à-dire qu'elle peut se voir associer une valeur de vérité (vrai ou faux). Elle est donc articulée autour d'un prédicat principal accompagné de ses actants (ceux-ci pouvant être des variables non-instanciées). Du point de vue de sa forme, la proposition élémentaire est écrite en pseudo-langue : une suite linéaire de « mots » français correspondant à des sémantèmes, qui respecte la syntaxe française mais où toute variation flexionnelle est évacuée. Par exemple :

(2) a. `1: X donner Y à Z`
    b. `6: X vraisemblable`

Toute proposition élémentaire commence par un numéro qui l'identifie de façon unique dans la définition. Ce numéro est séparé de la proposition proprement dite par les deux points suivis d'un espace. Formellement, une proposition élémentaire est équivalente à un sous-réseau sémantique dominé communicativement par le nœud correspondant au gouverneur syntaxique

de la proposition. Ainsi, la proposition (2a) est équivalente au sous-réseau ci-dessous (où le sens communicativement dominant est souligné) :




Les propositions élémentaires sont, comme les sous-réseaux sémantiques, destinées à se combiner pour former des représentations sémantiques complexes. La proposition élémentaire $P_2$ se combine avec la proposition $P_1$ lorsque le nœud dominant de $P_2$ est un actant du prédicat principal (ou, plus généralement, d'un prédicat) de $P_1$. Dans un tel cas, le numéro de référence de $P_2$ précédé d'un astérisque apparaît comme actant d'un prédicat dans la formule encodant $P_1$. Par exemple, l'ensemble de propositions (3) ci-dessous correspond à une structure sémantique où (i) le nœud dominant de la proposition `2` est le deuxième actant du nœud dominant de la proposition `1` et (ii) le nœud dominant de la proposition `3` est le deuxième actant du nœud dominant de la proposition `2`.

(3)
```
1: X communiquer *2 à Z
2: X consentir#1 à *3
3: X donner Y à Z
```

Nous laissons au lecteur le soin de construire le graphe sémantique correspondant à l'ensemble de propositions élémentaires ci-dessus. Celui-ci correspond à la composante centrale de la définition de CONSENTIR**2b** (DEC III). Notons que les numéros distinctifs de lexies apparaissent précédés d'un dièse (cf. `consentir#1` ci-dessus).

Comme nous l'avons dit, la proposition élémentaire comprend un prédicat avec tous ses actants. Un actant peut être soit une variable, soit un autre prédicat. Dans ce dernier cas, il est impératif que ce second prédicat ne gouverne pas un bloc de sens logiquement équivalent à une proposition : il ne peut gouverner un bloc syntaxique auquel serait associée une structure Thème ~ Rhème propre. Par exemple, dans `Y contraire à désir de X`, le prédicat ‹désir de X› n'est pas structuré par une opposition Thème ~ Rhème, comme le serait par exemple la formule `X avoir désir de Y`, qui correspond à une véritable proposition élémentaire de notre métalangage. Cet exemple illustre une propriété essentielle de l'encodage des propositions : il tient compte non seulement de la dominance communicative, mais aussi de l'opposition Thème ~ Rhème. Chaque sémantème se verra donc associer plusieurs encodages possibles, selon sa position dans la structure communicative de la proposition où il apparaît. Dans le cas de ‹désir›, nous avons ainsi rencontré les cinq possibilités d'encodage suivantes :

(4)
a. `désir de X de Y`
b. `désir de X`
c. `X avoir désir de Y`
d. `Y être désir de X`
e. `désir de X être Y`

Les éléments « lexicaux » de notre métalangage mentionnés jusqu'à présent dans nos exemples correspondaient tous soit à des sémantèmes du français soit à des prépositions vides permettant de connecter linéairement un prédicat à ses actants. Les propositions (4c-e) ci-dessus montrent que nous faisons aussi usage de verbes vides afin de modéliser la structure communicative dans l'encodage linéaire de la proposition. Ce recours aux verbes supports pour organiser les propositions élémentaires est lié au problème plus général de l'utilisation, dans notre langage formel, d'éléments lexicaux correspondant à des sens de fonctions lexicales. C'est un

sujet très intéressant, qui pose notamment le problème théorique de l'encodage des correspondants des fonctions lexicales dans les réseaux sémantiques Sens-Texte. Nous ne pouvons cependant pas l'aborder ici, faute de place.

### *3.1.2 Disjonctions de propositions élémentaires*

Il arrive fréquemment que les définitions du DEC contiennent des disjonctions de propositions. Par exemple, la proposition participiale ci-dessous est le présupposé donné dans le DEC IV pour la lexie DÉFIER**I.1** :

(5) *X croyant que Y a insulté X ou ⌜a porté atteinte⌝ à l'honneur de X*

Voici comment cette composante définitionnelle est analysée en propositions élémentaires dans la BDéf :

(6)
```
8: X croire#II *9
9.1: Y insulter X
9.2: Y _porter_atteinte_ à honneur de X
10: *9 passé
```

Nous pensons que cet encodage est élégant dans la mesure où il met en évidence la hiérarchie des différentes propositions impliquées de façon non ambiguë, tout en restant aisément lisible pour le lexicographe qui voudrait consulter directement les définitions de la BDéf. Ainsi, l'ensemble de propositions ci-dessus peut être lu en « métafrançais » de la façon suivante :

(7) *X croire que Y insulter X ou Y porter atteinte à honneur de X, dans le passé.*

Comme on le voit en (6) ci-dessus, les sémantèmes correspondant à des locutions du français sont encadrés par des traits de soulignement, traits qui lient aussi tous les éléments lexicaux constitutifs de la locution, par exemple : `_comme_si_`, `_ne_pas_`, `_porter_atteinte_`… Il nous arrive d'utiliser des unités sémantiques exprimées au moyen d'une combinaison de plusieurs éléments lexicaux mais qui ne sont pas associées à des locutions véritables. Dans de tels cas, seuls les traits internes de liaison sont utilisés, par exemple : `dans_le_but`, `pièce_fabriquée` (pour des parties d'artefact)…

Nous allons maintenant expliquer comment les propositions élémentaires sont regroupées dans les définitions de la BDéf. Nous devons pour cela introduire la notion de bloc définitionnel.

## 3.2 Bloc définitionnel

Il a souvent été mentionné dans la littérature Sens-Texte — voir, par exemple, (Polguère, 1997) et (Polguère, 2002) — que le sens d'une lexie ne formait pas un tout uniforme et qu'une bonne modélisation sémantique formelle devait posséder une structure interne explicite. En d'autres termes, la définition formelle doit mettre en évidence les différentes composantes d'une définition, composantes qui contribuent toutes à leur manière au sens global. Nos représentations indiquent explicitement les différentes composantes de sens, appelées *blocs définitionnels*. Au plus haut niveau de structuration, toutes les définitions se divisent en quatre blocs définitionnels majeurs : la composante centrale, les différences spécifiques, le typage des actants sémantiques et les relations sémantiques entre variables actancielles. Le second bloc majeur, les différences spécifiques, peut lui-même être subdivisé en blocs définitionnels élémentaires, correspondant chacun à une différence spécifique donnée, ou à un regroupement « naturel » de différences spécifiques (caractérisation de la forme, parties constitutives, fonc-

tions…). Nous allons brièvement décrire ces différents niveaux de structuration en utilisant des exemples tirés de la BDéf et en commençant par présenter un exemple complet de définition.

### *3.2.1 Présentation d'un exemple*

L'exemple qui va suivre va permettre au lecteur de se faire une idée générale de la nature formelle d'une définition de la BDéf, ce qui lui permettra de mieux saisir la suite de l'exposé. Il est clair que plusieurs éléments de formalisation présents dans la figure ci-dessous doivent être explicités. Nous espérons donc que le lecteur voudra bien se contenter pour l'instant d'une compréhension intuitive partielle de l'exemple en question. Nous avons sélectionné la définition de DÉFIER**I.1**, dont la version DEC originelle a déjà été présentée en (1) et qui a déjà été partiellement commentée dans son encodage BDéf.

```
1) Forme propositionnelle :
     X ~ Y en W à Z
2) Composante centrale :
     1: X communiquer à Y que *2
     2: X vouloir *3
     3: Y prendre_part à W avec Z
3) Différences spécifiques :
          /*But*/
     4: *1 dans_le_but *5
     5: X punir Y pour *9
          /*Moyen*/
     6: façon de X de *5 être *7
     7.1: X blesser#I1 Y
     7.2: X tuer Y
          /*Situation [présupposé]*/
     8: X croire *9
     9.1: Y insulter X
     9.2: Y _porter_atteinte_ à honneur de X
     10: *9 passé
4) Typage des variables :
     X: individu
     Y: individu
     Z: arme
     W: combat
5) Relations sémantiques entre variables actancielles :
     W[X,Y]
```

Figure 1 : Définition de DÉFIER**I.1** dans la BDéf

Une définition de BDéf est donc structurée en cinq blocs majeurs. Le premier bloc, la spécification de la forme propositionnelle, suit en tout point les conventions adoptées dans le DEC et n'appelle pas de commentaire spécial. Dans le cas particulier de la définition ci-dessus, la lexie vedette est décrite comme possédant quatre actants sémantiques. Les quatre autres blocs définitionnels majeurs vont maintenant être présentés.

### *3.2.2 Composante centrale de la définition*

La composante centrale est la paraphrase minimale de la lexie vedette. C'est la composante sur laquelle on s'appuie pour identifier l'étiquette sémantique associée à chaque lexie dans le DiCo ; sur la notion d'étiquette sémantique et de paraphrase minimale, voir (Polguère, 2003b). Dans le cas de DÉFIER**I.1**, on voit que notre modélisation identifie ‹X communique à Y qu'il veut que Y prenne part à W avec X› comme étant la paraphrase minimale de cette lexie. Cela apparaît très clairement dans la Figure 1 ; il est en revanche beaucoup plus difficile d'identifier

cette composante dans la définition linéaire originelle du DEC, déjà mentionnée en début d'article. Nous la répétons ici, avec la composante centrale indiquée en gras :

(8) *X défie Y en W à Z* = X croyant que Y a insulté X ou ⌜a porté atteinte⌝ à l'honneur de X, || **X communique à Y que X veut que Y prenne part dans** un combat **W** contre **X** en utilisant l'arme Z, dans le but de punir Y — pour ce que Y a fait à X — en blessant **I.1** Y ou en tuant Y.

La composante centrale est profondément enchâssée dans la séquence linéaire (8), « coincée » entre une composante présuppositionnelle fort complexe et d'autres différences spécifiques qui sont, elles, des éléments du posé. Comme nous allons le voir dans la section suivante, les définitions de la BDéf traitent explicitement les composantes présuppositionnelles comme des cas particuliers de différences spécifiques.

### *3.2.3 Différences spécifiques*

La définition de DÉFIER**I.1**, comme la plupart des définitions de la BDéf, voit son bloc définitionnel majeur de différences spécifiques subdivisé en sous-sections, c'est-à-dire en blocs définitionnels élémentaires. Le *rôle informationnel* de chaque bloc élémentaire est indiqué dans une en-tête apparaissant entre les marqueurs `/*…*/`. Par exemple, le bloc ci-dessous correspond à la caractérisation de la finalité du fait de défier**I.1** quelqu'un.

(9)
```
        /*But*/
4: *1 dans_le_but *5
5: X punir Y pour *9
```

Nous avons réalisé qu'il était important d'identifier clairement dans une en-tête le rôle informationnel de chaque composante plutôt que de se fier à la présence de tel ou tel sémantème (par exemple ici, ‹[dans le] but›) dans le bloc en question. En effet, le rôle informationnel n'est pas toujours explicitement indiqué dans une définition par une composante spécifique, le même rôle pouvant être modélisé de diverses façons en fonction des sémantèmes spécifiques devant être introduits dans la composante en question. La nécessité d'avoir une bonne paraphrase prime, dans les définitions analytiques du DEC, sur celle d'encoder explicitement le rôle informationnel de chaque composante.

La définition de la Figure 1 contient un bloc informationnel élémentaire présuppositionnel :

(10)
```
        /*Situation [présupposé]*/
8: X croire *9
9.1: Y insulter X
9.2: Y _porter_atteinte_ à honneur de X
10: *9 passé
```

La mention du statut communicatif présuppositionnel apparaît donc entre crochets, immédiatement après le nom du rôle informationnel. Cette convention peut être utilisée pour toute spécification communicative et pas simplement pour la présupposition.

Jusqu'à présent, nous avons procédé de façon en partie *ad hoc* pour nommer les différents rôles informationnels. Nous sommes cependant arrivés au stade où nous allons pouvoir, d'une part, rationaliser notre méthode d'encodage et, d'autre part, homogénéiser nos descriptions. Il est en effet possible d'isoler des patrons récurrents de structuration des différences spécifiques, patrons liés directement à l'étiquetage sémantique de la lexie vedette. Ainsi, toutes les lexies de type artefact peuvent potentiellement contenir dans leurs différences spécifiques les blocs définitionnels élémentaires suivants : `Fonction`, `Forme`, `Parties` et `Mode d'utili-`

`sation`. Les présupposés lexicaux sont, quant à eux, souvent associés aux blocs définitionnels `Caractéristiques`, `Situation`, `Source`, etc. Il est possible de mener une étude systématique de l'organisation interne des définitions, une fois celles-ci structurées en blocs définitionnels. Cette étude correspond à l'étape finale de notre recherche et va nous permettre de donner à la BDéf la cohérence et la rigueur formelle qui en fera un véritable outil de recherche sémantique.

### *3.2.4 Typage des actants*

Nous séparons le typage des actants sémantiques du reste des différences spécifiques. Dans la définition de DÉFIER**I.1**, nous avons ainsi[1] :

(11)
```
X: individu
Y: individu
Z: arme
W: combat
```

Ce choix s'explique par le fait que le typage des actants possède un statut informationnel très particulier. Un typage d'actant se comporte comme un présupposé s'appliquant dans la définition à chaque fois que la variable actancielle correspondante apparaît. Il « supervise » le fonctionnement de la définition et il serait arbitraire de l'introduire à un endroit plutôt qu'à un autre dans les différences spécifiques. Nous avons d'ailleurs noté un grand manque de cohérence dans la façon dont le typage est positionné dans les définitions du DEC. Il faudra mettre au point des conventions rigoureuses de typage lorsque nous utiliserons la BDéf, de façon inverse, comme source d'information pour générer des définitions linéaires « plus propres ».

### *3.2.5 Relations sémantiques entre variables actancielles*

Nous avons montré plus haut que les propositions élémentaires, grâce au formalisme utilisé, nous permettent d'encoder de façon non ambiguë les relations prédicat-actant, ce qui donne à notre encodage la puissance formelle d'un véritable réseau sémantique. Il arrive cependant fréquemment que des dépendances sémantiques unissent des variables actancielles dans les définitions. Il est donc nécessaire de stocker cette information de façon explicite, ce que nous faisons dans le dernier bloc définitionnel majeur. Dans le cas de DÉFIER**I.1**, la composante `W[X,Y]` indique que X est le premier actant de W et Y son second actant : ‹combat W entre X et Y›. C'est la position relative d'une variable dans la liste ordonnée entre crochets qui encode son numéro d'actant. Par exemple, la formule `Y[_,X]` servirait à indiquer que X est le second actant de Y, dont le premier actant n'apparaît pas explicitement dans la définition en question.

Remarquons, pour conclure cette section sur les blocs définitionnels, que les ponts sémantiques métaphoriques ou métonymiques entre acceptions d'un vocable sont indiqués explicitement sous forme de blocs définitionnels dans les différences spécifiques, avec comme en-tête de rôle informationnel `Lien_métaphorique`, parfois spécifié de la façon suivante : `Lien_métaphorique.Forme`, `Lien_métaphorique.Fonction`, etc. Seul ce type de pont sémantique est encodé explicitement dans les définitions des DEC publiés. Il faudra cependant, dans la BDéf, généraliser cette pratique à tous les ponts sémantiques, qu'ils soient ou non de nature métaphorique ou métonymique.

---

1. Notons que nous avons introduit un typage de X et Y qui n'apparaît pas dans la définition du DEC originelle. La construction de la BDéf est l'occasion de faire de nombreux ajustements dans le DEC.

### 3.3 Structure de la base de données de définitions

Du point de vue de son stockage informatique, la BDéf est gérée exactement de la même façon et avec les mêmes outils que le DiCo. C'est une base de données FileMaker, où chaque vocable est stocké sous forme d'un ensemble de fiches. Deux cas de figure se présentent : soit le vocable est monosémique, et sa description tient en une fiche ; soit le vocable est polysémique, et sa description tient en autant de fiches qu'il contient de lexies, plus une fiche additionnelle contenant des informations valides pour toutes les lexies du vocable (partie du discours, etc.).

Chaque fiche lexicale est structurée en autant de champs qu'il y a de types d'information à stocker. Ainsi, les cinq « sections » descriptives dont le nom apparaît en gras dans la Figure 1 ci-dessus sont stockées dans un champ particulier de fiche BDéf. Il faut ajouter aux cinq champs spécifiquement dédiés à l'encodage des définitions, des champs supplémentaires servent à stocker d'autres informations de nature lexicales ou propres au travail lexicographique : caractéristiques grammaticales de la lexie vedette, étiquette sémantique (pour établir une connexion sémantique avec les données du DiCo), exemples, statut d'avancement de la rédaction de la fiche, source (DEC I, II, III ou IV), auteur de la fiche et date de la dernière modification.

## 4 Étant d'avancement de la recherche et résultats à venir

À ce jour, la BDéf contient l'encodage des définitions de 310 vocables fortement polysémiques du DEC, soit 1040 définitions au total. Nous avons donc encodé environ deux tiers du DEC publié. Cette première étape du travail a été particulièrement longue puisqu'il a fallu tout d'abord mettre sur pied une stratégie et une méthode d'encodage, ce qui s'est effectué par tâtonnements successifs. Nous sommes maintenant entrés dans une phrase de révision des définitions encodées et de normalisation des en-têtes de blocs définitionnels. Nous comptons avoir fini d'encoder l'ensemble des définitions des DEC publiés d'ici la fin de notre projet de recherche, en mars 2004. Ce travail inclut la révision de toutes les définitions existantes ainsi que l'écriture de définitions supplémentaires visant notamment à compléter certains champs sémantiques, comme ceux des sentiments. Ce dernier travail pourra se faire en exploitant les données déjà réunies dans le cadre du projet DiCo-LAF (Polguère, 2000b).

Sans devoir attendre que le travail d'encodage soit complètement terminé, il sera possible d'entamer des recherches sur la sémantique lexicale du français exploitant les données contenues dans la BDéf. Nous comptons aussi utiliser le travail fait sur les définitions de DEC comme point de départ pour la confection d'un DEC électronique qui sera une version cumulative, révisée, étendue et entièrement informatisée des quatre volumes déjà publiés.

## Remerciements

Cette recherche est subventionnée par le projet 410-2001-0560 du Conseil de recherches en sciences humaines du Canada (CRSH). Les personnes suivantes sont également impliquées dans le projet BDéf : Lidija Iordanskaja, Suzanne Mantha, Igor Mel'čuk et Ophélie Tremblay. Nous remercions Ophélie Tremblay pour ses commentaires sur une première version de cet article.

## Bibliographie

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., Clas A., Polguère A. (1995), *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve, AUPELF-UREF/Duculot.

Milićević J. (2003), *Étude des aspects sémantiques et syntaxiques de la paraphrase : application à la génération automatique des phrases françaises*, Thèse de doctorat, Département de linguistique et de traduction, Université de Montréal.

Polguère, A. (1997), Meaning-Text Semantic Networks as a Formal Language. In L. Wanner (réd.) : *Recent Trends in Meaning-Text Theory*, Language Companion Series nº 39, Amsterdam/Philadelphia, John Benjamins, 1-24.

Polguère A. (1998), Pour un modèle stratifié de la lexicalisation en génération de texte, *Traitement Automatique des Langues* (*T.A.L.*), vol. 39, n˚ 2, 57-76.

Polguère A. (2000a), A "natural" lexicalization model for language generation, Actes de *Fourth Symposium on Natural Language Processing* (*SNLP'2000*), 37-50.

Polguère A. (2000b), Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French, Actes de *EURALEX'2000*, 517-527.

Polguère A. (2000c), Une base de données lexicales du français et ses applications possibles en didactique, *Revue de Linguistique et de Didactique des Langues* (*LIDIL*), n° 21, 75-97.

Polguère A. (2002), Le sens linguistique peut-il être visualisé ? In D. Lagorgette et P. Larrivée (réd.) : *Représentations du sens linguistique*, Lincom Studies in Theoretical Linguistics n° 25, Munich, Lincom Europa, 89-103.

Polguère A. (2003a), *Lexicologie et sémantique lexicale. Notions fondamentales*, Montréal, Presses de l'Université de Montréal, sous presse.

Polguère A. (2003b), Étiquetage sémantique des lexies dans la base de données DiCo, *T.A.L.*, à paraître.

Sérasset G., Polguère A. (1997), Outils pour lexicographes : application à la lexicologie explicative et combinatoire. Actes de *RIAO'97*, 701-708.

Wierzbicka A. (1985), *Lexicography and Conceptual Analysis*, Ann Arbor, Karoma Publishers.

Wierzbicka A. (1987), *English Speech Act Verbs. A semantic dictionary*, Sydney *et al.*, Academic Press.