# Présentation du *Dictionnaire Electronique des Mots* et de *Locutions Verbales* de Jean Dubois et Françoise Dubois-Charlier

Denis Le Pesant, Marie-Hélène Stéfanini
(1) MoDyCo, 200 avenue de la République 92001 Nanterre
(2) LIF, AMU, CNRS, 163 avenue de Luminy, 13288 Marseille Cedex 9
denis.lepesant@orange.fr, marie-helene.stefanini@lif.univ-mrs.fr

**Résumé.** Cet article est une présentation de deux ressources inédites de Jean Dubois et Françoise Dubois-Charlier : le *Dictionnaire Electronique des Mots* (DEM), base de données de plus de 140.000 entrées et *Locutions Verbales* (LOCVERB, 3.510 entrées). Ces deux ressources sont complémentaires de la base de données *Les Verbes Français* (LVF, 25.610 entrées). Après avoir évoqué LOCVERB dans ses relations avec LVF, nous décrirons le DEM. Une fusion ultérieure de ces trois ressources est envisagée.

**Abstract.** This article is a presentation of two new resources of Dubois Jean and Françoise Dubois-Charlier: the Electronic Dictionary of Words (DEM), a database of over 140,000 entries and Verbal Phrases (LOCVERB, 3510 entries). These two resources are complementary to the database French Verbs (LVF, 25,610 entries). After referring LOCVERB in its relations with LVF, we describe the DEM. Subsequent fusion of these three resources is considered.

**Mots clés :** Bases de données lexicales, Jean Dubois, Françoise Dubois-Charlier, dictionnaire des verbes, LVF, dictionnaire des locutions verbales, LOCVERB, dictionnaire des mots français, DEM.

**Keywords:** Lexical databases, Jean Dubois, Françoise Dubois-Charlier, dictionary of verbs, LVF, dictionary verbal phrases, LOCVERB, dictionary of French words, DEM.

## Introduction

Quelques jours avant l'Atelier que nous consacrons aux plus récentes ressources linguistiques de Jean Dubois et Françoise Dubois-Charlier, la dernière d'entre elles, le *Dictionnaire Electronique des Mots* (DEM), aura été publiée sous plusieurs formats sur les quatre sites suivants : le site FondamenTAL[1] du CNRS-LIF (Université Aix-Marseille), de CNRS-MoDyCo (Université Paris Ouest Nanterre), de RALI (Université de Montréal), et de NooJ.

Le projet d'une telle publication avait été déjà été évoqué dans (Dubois et Dubois-Charlier 2010), (Sabatier et Le Pesant 2013) et (Le Pesant, Sabatier, Silberztein, Stéfanini 2014). Elle fait suite à la publication en 2007 de la base de données lexicales des *Verbes Français* (LVF) sur les sites de CNRS-MoDyCo et de RALI (Université de Montréal), en parallèle avec celle d'un numéro de *Langue Française* (François, Le Pesant, Leeman 2007). Paul Sabatier a travaillé activement à la révision de toutes ces ressources. En même temps que le DEM, sera publiée une autre base de données lexicales des mêmes auteurs, intitulée *Locutions Verbales* (LOCVERB).

Après avoir évoqué LOCVERB dans ses relations avec LVF, nous ferons une présentation générale du DEM.

## 1 De *Verbes Français* (LVF) à *Locutions Verbales* (LOCVERB)

La base de données LOCVERB (locutions verbales) est totalement complémentaire de LVF (i.e les verbes simples) en ceci que les deux ressources ont exactement le même format. Bien que la révision de LOCVERB par Paul Sabatier ait

[1] http://www.talep.lif.univ-mrs.fr/Fondamental.html

été terminée il y a plus d'un an, cette ressource n'a pas encore pas été publiée. Comparons les deux bases de données lexicales.

Voici une représentation des cinq premières entrées de LVF (la version révisée par Paul Sabatier s'appelle « LVF + 1 »).

| | | | | | | |
|---|---|---|---|---|---|---|
| abaisser 01 | LOC | T3c | T1308 P3008 | r/d bas qc | baisser | On a~le rideau de fer,le store.Le rideau du magasin s'a~. |
| abaisser 02 | TEC | E3f | T13g0 P30g0 | f.ire qc VRS bas | incliner,pencher | On a~la manette,le levier.La manette s'a~vers le bas. |
| abaisser 03 | QUA | M3c | T1306 P3006 | f.mvt mns hauteur | baisser | On a~le mur d'un mètre.Le mur s'a~de beaucoup. |
| abaisser 04 | MON | M4b | T1306 P3006 | f.mvt mns val | baisser | On a~les prix,les revenus de dix p.c.Les prix s'a~de bcp. |
| abaisser 05 | MED | T4b | T1308 P3008 | r/d bas quant | faire descendre | Le malade a~la fièvre avec l'aspirine.La fièvre s'a~. |

TABLE 1 : LVF, les 5 premières entrées par ordre alphabétique des mots

Les utilisateurs de LVF y reconnaîtront les 7 rubriques principales, soit, de gauche à droite :

- rubrique MOT (ici 5 des 9 emplois du verbe *abaisser*)
- rubrique DOMAINE (locatif, technique, qualités, monnaie, écriture, médecine)
- rubrique CLASSE (Classe « générique » ; ex. « T3c » = verbes de transformation, sous-classe « 3 », subdivision « c »)
- rubrique CONSTRUCTION (« T1308 P3008 » = Verbe transitif à sujet inanimé concret + Emploi pronominal à sujet ; un ajout complément instrumental est fréquent, ce que marque le code « 8 »)
- rubrique OPERATEUR (codage de propriétés syntactico-sémantique ; « r/d bas qc » = « quelque chose est rendu ou devenu bas)
- rubrique SENS (synonyme ou parasynonyme)
- rubrique PHRASE (exemples illustrant le(s) emploi(s))

Pour « reconstruire » la classification syntactico-sémantique d'ensemble des verbes simples, il suffit d'opérer un tri sur (dans l'ordre) les rubriques CLASSE, CONSTRUCTION et MOT. C'est ce que représente la Table 2, où apparaît le début de la classe des verbes intransitifs de communication. Sur cet exemple figurent des indications de registre : par exemple « LANf » signifie « Domaine *Langage*, emploi *familier*). Il est à noter que le rôle syntactico-sémantique majeur est joué par le champ CONSTRUCTION.

| | | | | | | |
|---|---|---|---|---|---|---|
| bavarder 01 | LAN | C1a | A16 | loq mots ss cesse | causer bcp | On b~dans les couloirs. |
| causer 01 | LANf | C1a | A16 | loq mots | bavarder | On c~sans agir. |
| crier 03 | LAN | C1a | A16 | loq av force | hurler,forcer sa voix | On c~quand on appelle au secours. |
| écrire 05 | ECR | C1a | A16 | loq p écrit | s'exprimer ds écrit | On est en train d'é~. |
| écrivailler | ECRf | C1a | A16 | loq+ql p écrit | écrivasser | On é~dans de médiocres feuilles de chou. |

TABLE 2 : LVF, début des 2 premières sous-classes de Verbes de Communication

Rappelons (cf. (François, Le Pesant, Leeman 2007)) que LVF est subdivisé en 14 classes dites génériques : C (communication) ; D (donner) ; E (sortir, venir) ; F (frapper, toucher) ; H (états, comportements) ; L (localisation) ; M (mouvement) ; N (munir, démunir) ; P (sentiments, pensées) ; R (mettre en état le corps, fabriquer qqc) ; S (saisir, abandonner) , T (transformation) ; U (unir, détacher) ; X (auxiliaires, impersonnels, aspectuels).

Passons à LOCVERB (base de données de 3.510 entrées). Dans l'exemple suivant (Table 4), on reconnaît les mêmes 7 champs que dans LVF. Dans le champ CLASSE, « C4 » correspond à la sous-classe des verbes de communication figurant dans leur emploi « figuré » (cf. Table 3). Dans le champ CONSTRUCTION, « T1300 » signifie « verbe transitif direct à sujet humain et complément d'objet inanimé concret »; cette notation, qui peut surprendre s'agissant d'une locution permet de prédire une éventuelle transformation (ex. *Les cartes qui ont été abattues par Paul étaient pour le moins de mauvaise foi*).

| | | | | | | |
|---|---|---|---|---|---|---|
| abattre (ses) cartes | SOC | C4c | T1300 | ind qc caché abs | dévoiler son plan | On a~ses cartes devant P pour terminer la discussio |
| abattre (son) jeu | PSY | C4c | T1300 | ind qc caché abs | dévoiler son plan | On a~son jeu pour mettre un terme à la discussion. |
| abattre de la besogne | SOCf | H2c | A16 | f.travail | travailler bcp | On a~de la besogne dans cette ferme. |
| abattre du travail | SOC | H2c | A16 | f.travail | travailler bcp | On a~du travail dans cette ferme. |
| abonder dans le sens de | LIT | E2c | N1j | ire DS sens d | suivre qn | On a~dans le sens de l'orateur. |

TABLE 3 : LOCVERB, les 5 premières entrées par ordre alphabétique des mots

Le tri sur (dans l'ordre) les champs CLASSE, CONSTRUCTION et MOT appliqué à LOCV donne ceci, qui constitue les premières entrées de la première sous-classe des verbes de communication (locutions verbales intransitives) :

| | | | | | | |
|---|---|---|---|---|---|---|
| avoir (son) mot à dire | LAN | C1a | A16 | loq mots | ê en droit de parler | On a~son mot à dire dans cette négociation. |
| avoir des larmes dans la voix | VOX | C1a | A16 | loq mots émus | il parle d'une voix émue | Paul a des larmes dans la voix |
| avoir du coffre | LANf | C1a | A16 | loq mots fort | avoir de la voix | On a~du coffre et on peut brailler plus fort que toi. |
| avoir la langue bien affilée | VOXf | C1a | A16 | loq mots bcp | ê bavard, parler bcp | On a~la langue bien affilée et on amuse l'auditoire. |
| avoir la langue bien pendue | VOX | C1a | A16 | loq mots bcp | parler bcp,ê bavard | Le gardien a~la langue bien pendue. |

TABLE 4 : LOCVERB, début de la première sous-classe des Verbes de Communication

## 2 Le *Dictionnaire Electronique des Mots* (DEM)

Le DEM est une base de données de 145.333 entrées. Il réunit les entrées des deux autres dictionnaires électroniques de Jean Dubois et Françoise Dubois-Charlier qui viennent d'être évoqués. Il comprend en outre les mots appartenant à toutes les autres parties du discours, qu'il s'agisse de mots simples ou de mots locutionnels : noms, adjectifs, déterminants, adverbes, prépositions, conjonctions, interjections.

Le format du DEM n'est toutefois pas le même que celui de LVF et de LOCVERB. D'une certaine manière, le souci d'une *extension* maximale se fait au détriment d'un haut degré d'*intension.* Cela se manifeste par le fait que la base de données ne compte que 7 champs, soit le même nombre que pour LVF, mais pour plus de 6 fois plus de parties du discours. Du reste, les champs concernés ne correspondent que partiellement à ceux qui figurent dans LVF et LOCVERB.

Par ailleurs, il existe une propriété remarquable du DEM : il explicite les relations qu'il entretient avec LVF et LOCV, ce qui ouvre la voie à une éventuelle fusion des trois ressources, qui concilierait le souci d'une *extension* maximale avec celui d'un haut degré d'*intension.*

Voici une représentation des premières entrées du DEM, par ordre alphabétique des mots :

| M | CONT | DOM | OP | SENS | OP1 | CA |
|---|---|---|---|---|---|---|
| a 01 | tracer N | ECR | lett | alphabet latin | R3a1 | -1 |
| a 02 | artic N | PHN | voy | ouverte | C1a3 | -1 |
| à N (ê) | N rli qc à | RLA | st | (qc)appartenir à qn,qc | U3a1 | M- |
| à P inf | co str N | LIN | syn | ds le but de+inf | R4d1 | M- |
| à Pâques ou à la Trinité | adven adv | TPSm | st | jamais,à date incertain | L4a- | M- |
| à aucun moment | adven adv | TPS | st | jamais | L4a- | M- |
| à aucun prix | val adv | ECN | st | (céder)en aucun cas | H3f1 | M- |
| à bas N ! | excla P | VOX | intj | hostilité à N | C2d3 | R- |
| à bas prix | val adv | ECN | st | (vendre)à bon marché | H3f1 | M- |
| à base de N | fac adv | TEC | st | d composant principal | R3a1 | M- |

TABLE 5 : DEM, les premières entrées par ordre alphabétique des mots

Les rubriques du DEM sont :

- rubrique MOT
- rubrique CONTENU
- rubrique DOMAINE (ex. *écriture, phonétique, relation, linguistique, temps* etc.)
- rubrique OP(ERATEUR)
- rubrique SENS (synonyme, parasynonyme ou, parfois, définition)
- rubrique OP(ERATEUR)1
- rubrique CA(TEGORIE)

Les seuls champs communs aux dictionnaires de verbes et au DEM sont les champs MOT, DOMAINE et (en partie) SENS. S'agissant du champ CONTENU, voici ce qu'en disent les auteurs :

> « C'est l'articulation essentielle de ce dictionnaire, qui se veut syntaxique et syntagmatique. Chaque entrée est rangée, par CONT, dans une bulle/famille sémantico-syntaxique définie par le terme pivot de CONT. Exemples de termes pivots : *adhérer, advenir, alimentation* ou *vêtir*. Les différentes formules que l'on trouve pour une bulle donnée représentent les combinaisons syntaxiques avec le terme pivot.
>
> Par exemple : famille de CONT : *adhérer ;* formules : *adhérer à* N (ex : *christianisme*), *faire adhérer par N* (ex : *soviétisation*), *N q adhère* (ex : *marxiste*) ou *adhérer adjectif* (ex : *clérical*) ».

La rubrique OP(ERATEUR) « donne des précisions secondaires sur le référent du mot d'entrée, souvent en fonction/combinaison de ce qui est inscrit dans sa rubrique CONT. Par exemple, les mots *gifle* ou *tabasser* ont « frapper » comme terme pivot de CONT. La fonction de OP est de préciser s'il s'agit d'un *coup manuel*, de l'*utilisation d'une arme*, etc. ».

La rubrique OP(ERATEUR)1 est particulièrement remarquable en ceci qu'elle connecte le DEM avec LVF et LOCVERB (on y retrouve le système de codage des classes de LVF). En effet, cette rubrique « donne la classe de verbes, définie dans LVF, à laquelle le mot d'entrée est associé en vertu du terme pivot de sa rubrique CONT. Par exemple : *appartement, ville* ou *résider* ont « habiter » comme terme pivot de leur CONT. La classe correspondante dans LVF est L1a1 (= *être/se trouver quelque part* ) ». C'est aussi dans cette rubrique qu'on trouve des informations sur la formation des adjectifs (ex. « c » = adjectif non dérivé (*versatile*) ; « cn » = adjectif dérivé de nom (*poissonneux, tracassier*) ; « ca » adjectif dérivé d'adjectif (*vieillot*) ; « cvt » = adjectif dérivé de verbe transitif (*barbant*) » etc.).

La rubrique CA(TEGORIE) enfin, sur deux caractères, code la catégorie grammaticale et le genre du mot d'entrée. Par exemple « -1 », « M- » et « R- » (cf. Table 5 ci-dessus) codent respectivement « nom non-animé masculin », « adverbe » et « interjection »).

Cette présentation laisse imaginer quelle grande variété de requêtes (morphologiques, syntaxiques, sémantiques, ontologiques etc.) croisées ou non croisées est rendue possible à partir de cette énorme ressource syntactico-sémantique. Plusieurs publications ((Le Pesant et *alii* (à paraître) ; Sabatier et Le Pesant (2013)) montrent qu'on peut extraire du DEM des esquisses de véritables ontologies.

On reviendra sur ce point à plusieurs reprises au cours de notre Atelier. Qu'il nous suffise pour le moment d'évoquer un autre exemple que celui du domaine de la Musique, à savoir celui de l'Alimentation (Domaine « ALI»). Un tri portant (dans cet ordre) sur les champs DOM, CONT et OP permet d'obtenir d'excellents résultats dans la recherche d'une vue d'ensemble sur le vocabulaire du domaine. Se manifestent en effet successivement (par ordre alphabétique approximatif des CONT et en faisant abstraction de nombreux critères possibles de subdivision) :

- les adjectifs de qualité des aliments (ex. *aigre-doux, congelable*) ;
- les noms de préparation alimentaire autres que produits de la mer (ex.*andouille, beignet*) et de préparation à base de produits de la mer ; les noms de plats ;
- les noms d'outils de préparation des aliments (ex. *découenneuse*) ; les noms et verbes d'opérations culinaires diverses ;
- les noms de repas ; les verbes de manger ; les noms de mangeurs ;
- les noms d'entreprises de restauration et d'industries de l'alimentaire ;
- les noms de métiers de la cuisine, de la restauration et de l'industrie alimentaire.

## Conclusion

Cet ensemble de ressources a été d'ores et déjà implémenté par Max Silberztein grâce à la plate-forme d'ingénierie linguistique NooJ.

A l'horizon de ces travaux figure le projet de fusionner LVF, LOCVERB et DEM et de faire de cet ensemble un tout parfaitement cohérent, utilisable en TAL pour des tâches d'annotation syntaxique et sémantique et pour la création d'ontologies.

# Références

Dubois J., Dubois-Charlier F. (2010). La combinatoire lexico-syntaxique dans le *Dictionnaire électronique des mots*. Les termes du domaine de la musique à titre d'illustration. In Leeman D., Sabatier P. (ed). *Langages* 179-180, p.31-56.

François J., Le Pesant D. & Leeman D. (2007). Présentation de la classification des Verbes Français de J. Dubois et F. Dubois-Charlier. *Langue française* n°153 : 3-19.

Le Pesant, D., Sabatier, P., Silberztein, M., Stéfanini, M.-H. (sous presse). Présentation d'un thésaurus des mots d'affect : théorie, méthodes et applications. In Blumenthal, Novakova & Siepmann ed. *Nouvelles perspectives en sémantique lexicale et en organisation du discours. Actes du Colloque Emolex* (Osnabrück, 6-8 février 2013). Peter Lang pp. 395-406.

Le Pesant D., Sabatier P. (2013). Les dictionnaires électroniques de Jean Dubois et Françoise Dubois-Charlier et leur exploitation en TAL.In *Ressources Lexicales.* Gala N. et Zock M. ed. *Linguisticae Investigationes Supplementa* 30. Amsterdam : John Benjamins Publishing Company.

Leeman D., Sabatier P. éd. (2010). *Empirie, théorie, exploitation : le travail de Jean Dubois sur les verbes français*. *Langages* n°179-180.