Linx
Revue des linguistes de l'université Paris X Nanterre

80 | 2020
L'héritage de Jean Dubois et Françoise Dubois-Charlier

Deux dictionnaires informatisés de Jean Dubois et Françoise Dubois-Charlier, leurs ultimes travaux

Two computer-based dictionaries of Jean Dubois and Françoise Dubois-Charlier, their final works

Guy Lapalme et Denis Le Pesant

Édition électronique
URL : http://journals.openedition.org/linx/6671
DOI : 10.4000/linx.6671
ISSN : 2118-9692

Éditeur
Presses universitaires de Paris Nanterre

Référence électronique
Guy Lapalme et Denis Le Pesant, « Deux dictionnaires informatisés de Jean Dubois et Françoise Dubois-Charlier, leurs ultimes travaux », *Linx* [En ligne], 80 | 2020, mis en ligne le 10 juillet 2020, consulté le 05 août 2020. URL : http://journals.openedition.org/linx/6671 ; DOI : https://doi.org/10.4000/linx.6671

Département de Sciences du langage, Université Paris Ouest

Deux dictionnaires informatisés de Jean Dubois et Françoise Dubois-Charlier, leurs ultimes travaux

Two computer-based dictionaries of Jean Dubois and Françoise Dubois-Charlier, their final works

Guy Lapalme et Denis Le Pesant

Introduction

1 La plupart des dictionnaires dont Jean Dubois fut le maître d'œuvre chez l'éditeur Larousse sont des dictionnaires-papier traditionnels. Ce n'est que tardivement que Jean Dubois et Françoise Dubois-Charlier, dans la mouvance des dictionnaires informatisés du LADL[1], ont élaboré les dictionnaires informatisés que sont *Les Verbes Français* (désormais LVF) et le *Dictionnaire Electronique des Mots* (désormais DEM).

2 Compte-tenu du fait que LVF a déjà été commenté dans le numéro 153 de la revue *Langue Française* (François, Le Pesant et Leeman 2007) ainsi que dans les numéros 179-180 de la revue *Langages* (Leeman et Sabatier 2010), nous nous concentrerons moins sur la structure de la ressource elle-même que sur ses applications au TAL (Traitement Automatique des Langues). Nous nous étendrons en revanche davantage sur le DEM, qui est une ressource linguistique particulièrement mal connue et qui peut être considérée comme la synthèse des travaux lexicographiques de Jean Dubois et Françoise Dubois-Charlier. Certes, c'est un dictionnaire d'un formalisme peu commode et qui souffre d'être resté, à la mort de ses auteurs, dans un état d'inachèvement important, avec ce que cela implique de lacunes et de contradictions. Mais son extension peu commune et surtout ses corrélations avec LVF en font, selon nous, une source de données lexicales de premier ordre pour la linguistique du français et pour le TAL. Cet article est donc axé sur l'aspect lexicographique des ultimes travaux de Dubois & Dubois-Charlier : il s'agit d'une lexicographie informatisée disponible pour le TAL.

3 Dans la première partie nous évoquons les lexiques-grammaires du LADL, qui représentent le modèle dont s'inspirent Dubois et Dubois-Charlier pour LVF et le DEM.

4 Même si l'intérêt de mettre à la disposition de la recherche en TAL des ressources lexicales à la fois amples et raffinées est aujourd'hui souvent remis en cause, ces dernières restent à nos yeux indispensables. La publication du présent article aura été l'occasion, pour l'un de nous deux, Guy Lapalme, d'élaborer une version renouvelée, au format JSON, de l'interface de consultation des dictionnaires de Dubois et Dubois-Charlier qui figure sur un site dédié de l'Université de Montréal : http://rali.iro.umontreal.ca/rali/?q=fr/versions-informatisees-lvf-dem. Ce sera l'objet de la deuxième partie de cet article. Nous espérons de la sorte favoriser la diffusion de ces ressources lexicales auprès de la communauté des chercheurs en lexicologie, en lexicographie et en TAL.

5 La troisième partie de l'article se concentre sur le DEM. Notre conviction est que ce monument malheureusement inachevé, cette sorte de « lexique-grammaire total », ne devrait pas rester ignoré des linguistes et des spécialistes du TAL.

1. Des lexiques-grammaires du LADL aux dictionnaires informatisés de Jean Dubois et Françoise Dubois-Charlier

1.1 Les lexiques-grammaires du LADL

6 La notion de *lexique-grammaire* s'inscrit dans le contexte théorique des travaux de Zellig S. Harris qui a publié plusieurs versions de sa grammaire de l'anglais (*cf.* Daladier 1990 : 20). Une grammaire de Harris, telle Harris (1976), se présente comme un ensemble de règles organisé en un système axiomatisé. Mais il manque à ces règles leur *extension* lexicale. Le projet du *lexique-grammaire* élaboré par Maurice Gross consiste précisément en un programme d'explicitation de l'extension lexicale des règles. Prenons l'exemple d'une règle de la grammaire du français qui stipule (par définition *en compréhension*) que certains verbes, tel *étonner*, peuvent admettre en position sujet une proposition subordonnée de forme « (*le fait*) *que P* ». La définition *en extension* de cette « règle » ne sera rien d'autre que la liste exhaustive des verbes et locutions verbales qui admettent un sujet d'une telle forme, assortie d'une description détaillée des autres propriétés de chaque occurrence : c'est ce que réalisera Maurice Gross par exemple dans les Tables 4 et 5 (*cf. infra* Tableau 1) de *Méthodes en syntaxe* (Gross 1975 : 245-279).

7 Mais la notion de *lexique-grammaire* repose également sur un principe épistémologique très général, formulé ainsi par Maurice Gross :

> L'approche très répandue que nous critiquons et que nous nous proposons de remplacer, consiste essentiellement à effectuer des observations isolées dans des régions différentes de la structure linguistique. Ces observations ne sont jamais systématiques, et les « trous » qui les séparent ne sont pas explorés empiriquement. Les constructeurs de modèles laissent à l'induction le soin de combler ces trous. Cette foi dans l'induction ne repose sur aucune base scientifique (...). Toute construction théorique a toujours été précédée d'un long travail d'accumulation systématique des données et les chercheurs se sont toujours efforcés de combler les trous qui pouvaient se présenter dans leurs données avant d'avancer une règle générale (Gross 1975 : 9)

8 Dans la suite du texte, cette thèse est illustrée par l'exemple de ce que Kepler, pour la formulation des lois qui portent son nom, doit à l'accumulation de données astronomiques effectuée avant lui par Tycho Brahé.

9 D'un point de vue lexicographique, la théorie du lexique-grammaire se matérialise par l'élaboration par le LADL, dans les années 1970-1990, d'une série de bases de données informatisées. Il s'agit, pour des dictionnaires, d'un format révolutionnaire : au lieu d'être des textes, ce sont des tables presque immédiatement disponibles pour le TAL[2]. Chaque table représente une classe de mots partageant un ensemble de propriétés syntaxiques caractéristiques sous la forme d'une matrice binaire, à savoir un tableau rectangulaire de signes « plus » et « moins ». Les propriétés sont définies en haut de chaque colonne ; chaque entrée figure en ligne ; le reste de la ligne, c'est-à-dire l'ensemble des signes « + » et « – » constitue en quelque sorte la définition syntaxique de l'entrée. Voici en exemple un extrait de la Classe 4 de *Méthodes en syntaxe* (Gross 1975) :

Tableau 1 : Version simplifiée de la Table 4 des verbes (V_4lgt.xls) du lexique-grammaire[3], diffusée sur le site *Infoling* de l'Université de Marne-la-Vallée et correspondant à la Table 4 de M. Gross (1975))

<ID>	N0 =: Nhum	N0 =: le fait Qu P	N0 =: V1-inf W	<ENT>Ppv	Ppv =: Neg	<ENT>V	Neg	N1 =: le fait Qu P	N1 se V de ce Qu P	N1 se V auprès de N3hum de ce Qu P	N1 être Vpp de ce Qu P	[passif par]	[passif de]	N0 V N1 contre N2hum	<OPT>
1	+	+	+	<E>	-	abasourdir	-	-	-	-	+	+	-	-	Cette nouvelle a abasourdi Max
2	+	+	+	<E>	-	abattre	-	-	+	-	+	+	-	-	Que Paul ait dit cela a abattu Max
3	+	+	+	<E>	-	abêtir	-	-	+	-	+	+	-	-	Que Paul soit toujours là abêtit Maria
4	+	+	+	<E>	-	abîmer	-	-	-	-	-	+	-	-	Que Paul ait révélé cela a abîmé Marie
5	~	~	~	<E>	-	abrutir	-	~	~	~	~	~	~	~	Que Max ait tant parlé a abruti Luc
6	+	+	+	<E>	-	abuser	-	-	-	-	-	+	-	-	Que Luc ait fait triché a abusé Max
7	+	+	+	<E>	-	accaparer	-	-	-	-	-	+	-	-	Ces tâches accaparent Léa
8	+	+	+	<E>	-	accomplir	-	-	+	-	-	+	-	-	Le travail a accompli Max
9	~	~	~	<E>	-	accrassiner	-	~	~	~	~	~	~	~	Cette situation a accrassiné Marie
10	+	+	+	<E>	-	accrocher	-	-	-	-	-	+	-	-	Ce spectacle a accroché Marie

10 Il est à noter que l'aspect sémantique des travaux du LADL est très important. Ainsi la Table 4 du Lexique-grammaire est-elle présentée par Maurice Gross de la façon suivante :

> Les verbes de cette table sont sémantiquement homogènes. La grande majorité d'entre eux correspond à un sentiment « déclenché » par N0 et « éprouvé » par N1. Les principales propriétés représentées sont celles de la complétive sujet, ainsi que celles de la forme adjectivale associée *N0 est V-a pour N1* (M. Gross 1975 : 170)

11 Le Lexique-Grammaire du LADL[4], dans cette version « modernisée » élaborée par Elsa Tolone, est diffusé au format CSV, sous licence LGPL-LR (version 3.4)[5].

12 Ont été successivement élaborées au LADL, outre celles dont il vient d'être question, les tables de constructions intransitives (Boons, Guillet, Leclère 1976) ; les tables de constructions locatives (Guillet, Leclère 1992). Beaucoup d'autres travaux ont été effectués par les membres du LADL sur les adjectifs et les noms prédicatifs. Des tables de plusieurs milliers d'expressions figées ont été également constituées[6].

13 Au début des années 1990, Maurice Gross confie à Max Silberztein la tâche de construire le logiciel d'ingénierie linguistique INTEX. Le système compile et lemmatise toute classe de mots formalisée en matrice binaire de traits syntactico-sémantiques. Le format des grammaires locales INTEX est celui d'automates finis (*cf.* Perrin, 1989)

corrélés à des dictionnaires (compilés dans le format d'INTEX) pour des applications de traitement informatique de corpus textuels.

14 En TAL, les deux défis pour la recherche d'informations dans des corpus textuels sont le phénomène de la polysémie et celui de la polylexicalité (mots composés, locutions, expressions à verbe support et phrases figées). Le LADL aura eu le mérite d'avoir démontré par des expériences empiriques approfondies que ces deux phénomènes, loin d'être marginaux, sont massifs et font partie des propriétés les plus remarquables des langues naturelles. Son autre mérite aura été d'avoir conçu des outils informatiques pour relever ces défis.

15 Jean Dubois et Françoise Dubois-Charlier, dans leurs derniers travaux, se conformeront pour l'essentiel à la méthode ayant présidé à l'élaboration des lexiques-grammaires du LADL.

1.2 Généralités sur *Les verbes français* (LVF) de Dubois & Dubois-Charlier

16 Les travaux préparatoires à l'élaboration de LVF ont eu lieu dans les locaux du LADL, à l'Université de Paris-Jussieu. Dans les années 1980, se tenait là une réunion tous les mardis sur la classification des verbes. Elle était présidée par Jean Dubois, et beaucoup de chercheurs du LADL y participaient, tout particulièrement Alain Guillet, qui joua, avec Françoise Dubois-Charlier, un rôle très important dans cette entreprise colossale. La base de données, qui ne fut rendue publique que vers 2010, servit de base à une version papier publiée par Larousse (Dubois & Dubois-Charlier 1997). Au début des années 2010 fut publiée une version informatisée de LVF, dite « LVF+1 », revue et corrigée à partir de l'original par Paul Sabatier et sous la supervision des deux auteurs.

17 La principale différence de LVF avec les dictionnaires du LADL réside dans le fait que les auteurs ont organisé leur classification selon 14 classes syntactico-sémantiques :

Tableau 2 : les 14 classes syntactico sémantiques de LVF

Code	Classe		Code	Classe
C	*Communication*		**N**	*Munir, démunir*
D	*Don, privation*		**P**	*Psychologie*
E	*Entrée, sortie*		**R**	*Mise dans tel ou tel état*
F	*Frapper, toucher*		**S**	*Saisir, posséder*
H	*Etat physique*		**T**	*Transformation*
L	*Localisation*		**U**	*Union, réunion*
M	*Mouvement*		**X**	*Verbes auxiliaires*

18 L'autre différence entre LVF et les lexiques-grammaires est que la table, unique (elle comporte 25 600 lignes), n'a pas la forme d'une matrice de traits binaires. Les colonnes énoncent, soit de façon explicite, soit de façon codée, un ensemble de propriétés morphologiques, syntaxiques et sémantiques. Voici une image des premières entrées (par ordre alphabétique) de LVF +1, sur tableur Excel :

Tableau 3 : Extrait de LVF + 1, tri par ordre alphabétique des entrées (version abrégée)[7]

	DOMAINE	OPERATEUR	CLASSE	SENS	PHRASE	CONSTRUCTION	A. table
abaisser 01	LOC	(#) [r/d] bas qc	T3c	baisser	On a~ le rideau de fer, le store. Le rideau du magasin s'a~.	T1308 - P3008	1
abaisser 02	TEC	(#) [f.ire] qc VERS bas	E3f.3	incliner, pencher	On a~ la manette, le levier. La manette s'a~ vers le bas.	T13g0 - P30g0	-
abaisser 03	QUA	(#) [f.mvt] moins hauteur	M3c	baisser	On a~ le mur d'un mètre. Le mur s'a~ de beaucoup.	T1306 - P3006	-
abaisser 04	MON	(#) [f.mvt] moins valeur	M4b.1	baisser	On a~ les prix, les revenus de dix pour cent. Les prix s'a~ de beaucoup.	T1306 - P3006	-
abaisser 05	MED	(#) [r/d] bas quantité	T4b	faire descendre	Le malade a~ la fièvre avec l'aspirine. La fièvre s'a~.	T1308 - P3008	-
abaisser 06	PSYt	(#) [m.e.état] mauvais qn abstrait	R2a.3	avilir, humilier	On a~ P, son orgueil en le blâmant. On s'a~ en public.	T1108 - P1000	-
abaisser 07 (s')	PSY	(#) [f.mvt] humble A+infinitif	M2b.2	s'humilier à	On s'a~ à demander une faveur, à cette demande.	P10a0	-
abaisser 08 (s')	VEH	(voie) [lc.qp] vers bas	L3a.1	descendre vers	La route s'a~ vers la rivière.	P3001	-
abaisser 09 (s')	PSY	(#) [f.mvt] A état vil	M2b.2	s'avilir, tomber jusqu'à	On s'a~ au niveau de cet escroc.	P10a0 - T11a0	-
abalourdir (s')	PSYt	(#) [r/d] balourd (adjectif)	T2b	abêtir, abrutir	On s'a~ avec un tel film. On a~ P avec ce livre. La télé a~ P.	P1000 - T9106	-
abandonner 01	DRO	(#) [dat] qc A qn	D2a	laisser, léguer	On a~ ses biens à ses enfants, à une fondation.	T13a0	-
abandonner 02	MAR	(#) [dat] mouvement A qc	D3a.2	laisser aller	On a~ sa barque à un fort courant.	T13a8 - P30a8	-
abandonner 03	EQU	(#) [dgrp] rênes	S3b.1	lâcher, laisser aller	On a~ les rênes de l'attelage.	T1300	-
abandonner 04	PSY	(#) [dgrp] abstrait	S4b.1	renoncer à	On a~ un projet, une idée. Le projet est a~.	T1300	-
abandonner 05	SOC	(#) [dgrp] abstrait	S4b.1	quitter, lâcher	On a~ ses études. On n'a~ pas après un premier échec.	T1300 - A10	-
abandonner 06	SPO	(#) [dgrp] lutte	S2e.1	s'avouer vaincu	On a~ le match. Le boxeur a~ au premier round.	T1300 - A10	-
abandonner 07	COM	(#) [li.mut] CONTRE/POUR	T3f.2	quitter	On a~ un appartement pour une maison.	T13k0	-
abandonner 08	SOC	(#) [dgrp] lieu	S3h	quitter, inhabité	On a~ ce village devenu désert. Le village est a~.	T1307	-
abandonner 09	LOC	(#) [dgrp] qn qp	S2d.1	quitter, laisser	On a~ un chien sur la route, un enfant dans la voiture.	T1101	-
abandonner 10	SOC	(#) [dgrp] abstrait	S4b.1	renoncer à	On a~ ses responsabilités. Après cet échec, on a~.	T1300 - A10	-

19 Considérons la première entrée du Tableau 3, *abaisser 01*. Le champ « Construction » *T1308-P3008*, code les propriétés suivantes : verbe transitif direct à sujet_humain susceptible d'avoir un ajout instrumental (« T1308 »), ainsi que d'être conjugué à la voix pronominale, avec dans ce cas un sujet non-humain (« P3008 »). Quant au champ « Classe », il renvoie à une des sous-classes des 14 classes syntactico-sémantiques évoquées *supra*. La classe T regroupe les verbes de « transformation » ; la sous-classe T3c est définie par la Construction *T1308-P3008* ; le champ « Opérateur » indique que le verbe *abaisser* 01 a le même sémantisme que la valeur causative de l'adjectif de qualité physique *bas* ; et un des champs morphologiques renvoie au dérivé *abaissable*. Enfin, le champ « Domaine » indique que ce verbe appartient au domaine de la localisation (« LOC »).

20 Cette rapide présentation donne, nous l'espérons, une idée des opportunités qu'offre le dictionnaire informatisé LVF en matière de tris croisés selon un grand nombre de critères linguistiques.

1.3 Généralités sur le Dictionnaire Electronique des Mots (DEM)

21 Il nous faut répéter que le *Dictionnaire Electronique des Mots* (DEM) est une œuvre très largement inachevée, moins du point de vue de l'extension (145 197 entrées) que du point de vue de la cohésion interne et de la révision des détails. Les auteurs de ces lignes avaient demandé à Françoise Dubois-Charlier, déjà souffrante de la maladie qui allait l'emporter quelques mois plus tard, la permission de publier une version XML du DEM. Elle n'a accepté qu'avec les plus grandes réticences, arguant du caractère imparfait du dictionnaire. Les auteurs du DEM n'auront pas eu le temps de rédiger un manuel d'utilisation ; c'eût été indispensable, compte tenu du formalisme complexe et

peu intuitif qui caractérise ce travail. En revanche, ils avaient eu l'occasion de publier une présentation du DEM dans un article de revue (*cf.* Dubois et Dubois-Charlier 2010).

22 Une des plus intéressantes propriétés du DEM est sa connexion explicite avec LVF. Cela va nous permettre de présenter ici (*cf. infra* Section 2.2) une nouvelle interface de consultation commune aux deux ressources.

23 Soit cet extrait du DEM :

Tableau 4 : Extrait du DEM

MOT	**CONT**	**DOM**	**OP**	**SENS**	**OP1**	Genre
carnau	fac v N	BAT	ins	conduit d cheminée	R3a1	1
moise	fac v N	BAT	ins	couple 2 pièces jumelles	R3a1	2
escalier	fac v N	BAT	ins	marches pr monter	R3a1	1
escalier de service	fac v N	BAT	ins	réservé au personnel	R3a1	1

24 Les quatre entrées mettent en relation quelques noms d'artefacts dans le domaine du *Bâtiment* avec dans le champ « OP1 » la mention de la classe R3a1, qui est une sous-classe définie ainsi dans LVF : « fabriquer quelque chose de concret avec un appareil » ; les entrées concernées de LVF sont *bâtir*, *construire* (*construction*), *édifier* (*édification*), *fabriquer* (*fabrication*), *maçonner*.

25 Avant d'y revenir plus longuement, contentons-nous pour le moment d'indiquer que le DEM a pour matière la multitude des fichiers lexicographiques dont disposait Jean Dubois en tant que principal collaborateur, pendant 30 ans, du directeur de la rédaction des dictionnaires Larousse, son frère Claude Dubois. D'autre part, Jean Dubois et Françoise Dubois-Charlier ont travaillé sur le DEM dès le milieu des années 1990, une fois achevé leur précédent projet (LVF), et jusqu'à quelques mois de leurs morts respectives, en 2015 et 2016.

2. Nécessité d'une large diffusion des dictionnaires de Dubois & Dubois-Charlier auprès de la communauté des linguistes et des chercheurs en TAL

2.1 Ressources lexicales et TAL contemporain

26 Tous les systèmes de TAL sont basés sur des données linguistiques et de représentations en fonction des objectifs poursuivis : par exemple, la recherche et l'extraction d'informations dans des textes, la traduction, la correction, la génération de texte, etc.

27 Depuis les premiers travaux en traduction automatique il y a près de 50 ans, pratiquement tout le travail en TAL s'est appuyé sur des ressources fines de la langue : dictionnaires qui recensaient l'ensemble des mots d'une langue, des règles de conjugaison et déclinaison ainsi que des grammaires pour établir des relations entre les

mots. Les systèmes ont pu s'appuyer sur une longue tradition de travaux en linguistique tant formelle que basée sur des corpus.

28 En français, on peut penser, en-dehors des travaux évoqués précédemment de Maurice Gross et de son équipe du LADL, dont faisait partie Morris Salkoff (*cf.* la grammaire en chaîne, Salkoff 1980), à ceux de K. van den Eynde et P. Mertens (2006) ou encore d'Igor Mel'čuk (1997) et son équipe du *Dictionnaire explicatif et combinatoire*. Les ressources issues de ces projets ont fait l'objet de formatages, de mises au point, d'enrichissements et d'exploitations dans la communauté TAL.

29 Aujourd'hui en 2020, certains remettent en cause la pertinence ou l'actualité de ce type de ressource, la tendance récente en TAL étant basée sur l'analyse de corpus pour transformer les mots en vecteurs (Mikolov *et al.* 2013, Devlin *et al.* 2019) sur lesquels on effectue des calculs pour établir des liens entre les mots ou groupes de mots, pour déterminer le type ou la polarité d'un texte, pour détecter si le texte est un pourriel ou non, etc. Ces calculs sont basés sur des paramètres déterminés par le traitement de grandes quantités de texte. Or au moins une partie de ces textes doit toutefois avoir été préalablement annotée plus ou moins finement en fonction de l'application visée : par exemple, en parties du discours pour l'analyse lexicale, avec des liens entre mots pour l'analyse syntaxique, en genres de texte pour la classification, etc.

30 Pour certaines applications très particulières où l'entrée et la sortie du système sont du texte, par exemple la traduction (Johnson *et al.* 2017) ou la reconnaissance de la parole (Wang *et al.* 2019), il est maintenant possible de développer des systèmes *de bout en bout* (*end-to-end*) où des algorithmes réussissent à établir des correspondances entre des éléments de l'entrée et de la sortie qu'ils peuvent reproduire sur de nouveaux textes, sans avoir besoin de créer des niveaux intermédiaires de représentations symboliques, comme le faisaient les travaux précurseurs, au temps où la vitesse de calcul et la mémoire des ordinateurs n'auraient pas permis ce type de traitements qu'on prend maintenant pour acquis. Pour y arriver, il « *suffit* » de recueillir de grandes quantités de textes ou des pistes sonores avec les sorties correspondantes, ce qui n'est pas toujours facile, mais des techniques efficaces ont été développées pour tirer parti de grandes quantités de textes maintenant disponibles en format informatisé.

31 On est donc loin des travaux des linguistes *traditionnels* qui travaillaient sur des exemples souvent artificiels ou créés de toute pièce à partir de leurs intuitions ou de leur connaissance fine de la langue. Il faut aussi souligner que les chercheurs en TAL ne cherchent plus toujours la *perfection;* ils mesurent plutôt leur succès en termes de mesures de précision et de rappel sur des tâches standardisées (Paroubek *et al.* 2007).

32 Pour la plupart des autres tâches, où la sortie n'est plus du texte, mais plutôt un jugement (est-ce un pourriel ou non ? le texte exprime-t-il un sentiment positif ou négatif ? trouver une réponse à une question demandant un certain raisonnement, qui a fait quoi, à qui, comment, quand et où ?, etc.), les systèmes TAL doivent s'appuyer sur des corpus et des ressources compilées au fil des années par des linguistes, telles WordNet (Miller 1995) ou FrameNet (Ruppenhofer *et al.* 2016). Les systèmes s'appuient aussi sur des ressources collaboratives comme Freebase, Wikipedia ou Wiktionnaire. Dans certains domaines, on tire aussi parti d'ontologies créées pour d'autres fins, par exemple SNOMED dans le domaine médical. Ce sont toutes des ressources créées manuellement au fil des années par des équipes de chercheurs.

33 Cela démontre bien l'intérêt de ces ressources lexicales qui profitent maintenant à des générations de chercheurs toujours en quête de données annotées de façon

systématique. Même si des méthodes d'apprentissage non supervisé ont été développées, il est toujours nécessaire d'amorcer la pompe avec des données validées. Des données annotées sont également essentielles afin d'évaluer les résultats des systèmes.

34 Aucune ressource lexicale n'est complète, mais combinée à d'autres, elle peut être utile pour aider à ajouter des informations. Souvent, seul un sous-ensemble de la ressource n'est utilisé pour une application. Un bon exemple est WordNet (Miller 1995), développé originellement comme un modèle psychologique de l'accès lexical via des liens de synonymie. Il n'est jamais utilisé en totalité : dans certains cas, ce sont les ensembles de synonymes, dans d'autres cas c'est la liste des entrées, les définitions ou les regroupements via les liens d'hyperonymie. Des thésaurus compilés au cours des siècles derniers (*cf.* Lewis 1978) ont aussi servi dans plusieurs applications, notamment pour la désambiguïsation dans le cadre de systèmes de question-réponse.

35 Rappelons les caractéristiques des ressources des Dubois et Dubois-Charlier:

- *Les verbes français* (LVF) : plus de 25 000 entrées correspondant à plus de 12 000 verbes différents répartis en classes et schèmes d'utilisation ;
- *Dictionnaire électronique des mots* (DEM) : plus de 145 000 mots et expressions françaises codifiées permettant de les regrouper selon certains critères.

36 LVF et DEM ont l'avantage d'avoir été compilés par deux linguistes de renom avec un souci d'uniformité qu'on retrouve rarement dans les ressources lexicales actuelles. Elles sont beaucoup plus qu'une simple nomenclature de mots ou expressions, ce sont avant tout des regroupements selon des critères syntaxiques et sémantiques. Ces ressources comportent évidemment des lacunes, mais il nous semble intéressant de les considérer dans leur ensemble et de tenter d'en profiter pour combiner ces informations avec les différents domaines du TAL.

37 Des utilisations de ces ressources pour le TAL ont été présentées lors de communications dans des colloques, comme celles de Bédaride (2012), de Danlos, Nakamura et Pradet (2014), de Guillaume, Fort, Perrier et Bédaride (2014), de Mazziotta (2014). Quelques articles ont été également publiés, notamment ceux de Silberztein (2010) et de Sabatier et Le Pesant (2013). En-dehors du domaine du TAL, les dictionnaires informatisés de Dubois & Dubois-Charlier ont fait l'objet de plusieurs études dans des revues de linguistique française, notamment celles de François, Le Pesant et Leeman (2007), Dutoit et François (2007), Leeman et Sabatier (2010) et François (2017).

2.2 Les interfaces de consultation

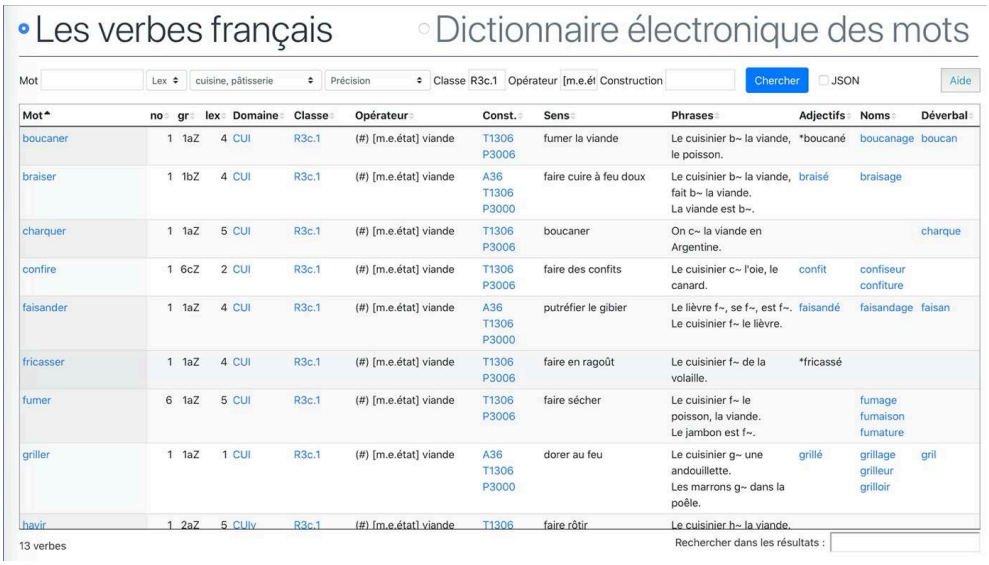
38 Depuis 2008, nous avons publié une série de versions du LVF et du DEM sous plusieurs formats différents (Excel, CSV, XML), sur les sites de l'Université de Paris Nanterre pour les versions Excel et de l'Université de Montréal pour les versions Excel, CSV, XML. Au cours des dernières années, JSON est devenu le langage de balisage le plus utilisé dans le domaine du TAL, pour lequel plusieurs outils de manipulation très efficaces ont été développés. Nous avons donc décidé, à l'occasion de la présente publication, de développer une version JSON[8] de ces mêmes ressources en espérant qu'elles pourront stimuler leur diffusion et leur utilisation. JSON est un moyen de codifier une structure de données sous forme de champs étiquetés. Contrairement à une base de données relationnelle, chaque enregistrement peut être plus ou moins détaillé, ce qui

correspond bien à la situation d'un dictionnaire où certaines entrées sont très élaborées alors que d'autres ne comportent que quelques informations. Cette version JSON est une réorganisation et une augmentation, décrite à la prochaine section, par rapport à la version XML que nous avons publiée il y a quelques années.

39 Les deux ressources sont diffusées sous la forme d'expressions JSON, chaque entrée occupant une ligne (un format souvent qualifié de *JSON Lines*[9]) : chaque entrée du lexique occupe une ligne du fichier, chaque ligne étant une expression JSON, elle peut être traitée facilement avec toute librairie manipulant ce type de données. Ces extractions peuvent ensuite être manipulées facilement avec d'autres outils bien connus des informaticiens-linguistes.

40 Cette version nous a permis de développer une nouvelle interface de consultation *intégrée* pour LVF et le DEM[10]. Les Figures 1 et 2 ci-dessous donnent des exemples de requêtes combinées autour d'un même domaine sémantique auprès de LVF et du DEM. Après avoir choisi la ressource dans la première ligne, il suffit d'entrer quelques contraintes dans des champs de texte (dans la deuxième ligne) pour faire afficher un tableau donnant les entrées qui respectent toutes ces contraintes. En plus de faciliter l'exploration des ressources séparément soit par filtrage ou regroupement, l'interface permet de suivre des liens entre le LVF et le DEM, car les auteurs avaient déjà imaginé de tels liens. Par exemple, dans le DEM, le champ OP1 fait référence à un schème du LVF.

41 Imaginons qu'on ait besoin, à partir des deux ressources, de rechercher les termes relevant du domaine de la cuisine : certains sont des verbes de préparation des aliments à des fins culinaires qu'on trouvera dans LVF (ex. *cuire*), les autres des noms d'aliments, qu'on trouvera dans le DEM (ex. *viande, légume*). Supposons d'autre part qu'on veuille se limiter au cas des aliments carnés. L'interface intégrée de consultation de LVF et du DEM permet qu'on se persuade, par une série de sondages sur les mots isolés, que dans LVF c'est la corrélation du *Domaine cuisine, pâtisserie* avec la Classe « *R3c.1* » qui donne accès aux verbes d'opérations culinaires. On découvre aussi qu'en utilisant le champ *Opérateur* on a accès à des sous-domaines : il en existe un de forme « [m.en état] viande ». Le tri selon ces trois critères a pour résultat une liste de 13 verbes et une liste de 15 noms partageant le même radical qu'un verbe de la liste :



Mot	no	gr	lex	Domaine	Classe	Opérateur	Const.	Sens	Phrases	Adjectifs	Noms	Déverbal
boucaner	1	1aZ	4	CUI	R3c.1	(#) [m.e.état] viande	T1306 P3006	fumer la viande	Le cuisinier b~ la viande, le poisson.	*boucané	boucanage	boucan
braiser	1	1bZ	4	CUI	R3c.1	(#) [m.e.état] viande	A36 T1306 P3000	faire cuire à feu doux	Le cuisinier b~ la viande, fait b~ la viande. La viande est b~.	braisé	braisage	
charquer	1	1aZ	5	CUI	R3c.1	(#) [m.e.état] viande	T1306 P3006	boucaner	On c~ la viande en Argentine.			charque
confire	1	6cZ	2	CUI	R3c.1	(#) [m.e.état] viande	T1306 P3006	faire des confits	Le cuisinier c~ l'oie, le canard.	confit	confiseur confiture	
faisander	1	1aZ	4	CUI	R3c.1	(#) [m.e.état] viande	A36 T1306 P3000	putréfier le gibier	Le lièvre f~, se f~, est f~. Le cuisinier f~ le lièvre.	faisandé	faisandage	faisan
fricasser	1	1aZ	4	CUI	R3c.1	(#) [m.e.état] viande	T1306 P3006	faire en ragoût	Le cuisinier f~ de la volaille.	*fricassé		
fumer	6	1aZ	5	CUI	R3c.1	(#) [m.e.état] viande	T1306 P3006	faire sécher	Le cuisinier f~ le poisson, la viande. Le jambon est f~.		fumage fumaison fumature	
griller	1	1aZ	1	CUI	R3c.1	(#) [m.e.état] viande	A36 T1306 P3000	dorer au feu	Le cuisinier g~ une andouillette. Les marrons g~ dans la poêle.	grillé	grillage grilleur grilloir	gril
havir	1	2aZ	5	CUIv	R3c.1	(#) [m.e.état] viande	T1306	faire rôtir	Le cuisinier h~ la viande.			

Figure 1 : Requête sur LVF selon les critères du Domaine, de la Classe et de l'Opérateur

42 Pour trouver les noms d'aliments susceptibles de se combiner aux verbes, on interroge le DEM. En procédant une nouvelle fois par sondage sur des mots individuels, on se persuade rapidement qu'il faut sélectionner dans le champ CONT le symbole « *ali N* » en corrélation avec le *Domaine* de la *boucherie*. On obtient une liste de 70 noms de viandes, dont deux locutions. En limitant les résultats à ceux qui contiennent le mot « bœuf », on obtient les 25 mots (*araignée, bronne, bavette* etc.) apparaissent dans la Figure 2. Il est à noter que le champ OP1, qui met le DEM en relation avec LVF, renvoie non pas aux termes de préparation culinaire (ce qui serait redondant), mais à la Classe *S3j1* qui rassemble des verbes de consommation alimentaire (*avaler, consommer, déglutir*, etc.).

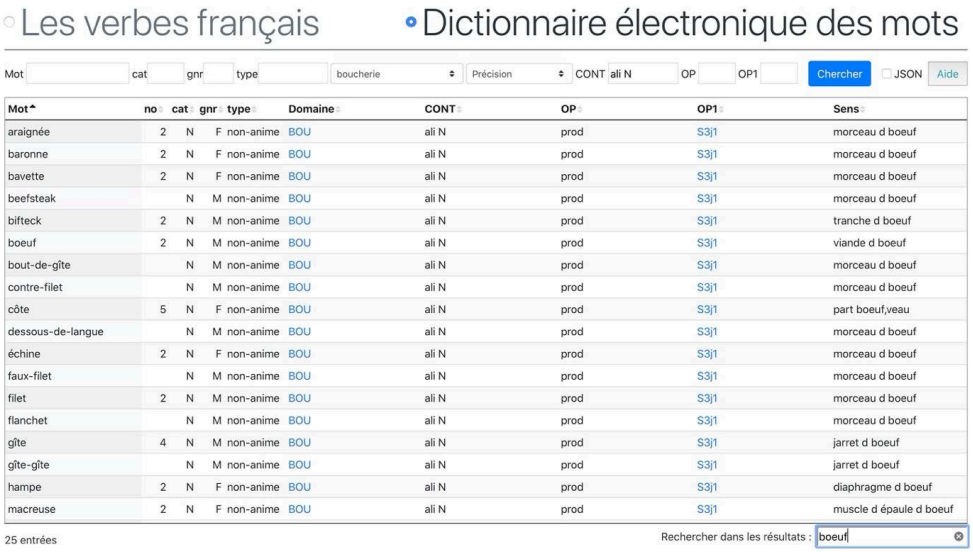


Figure 2 : Requête sur le DEM selon les critères du Domaine, du Contenu et de la présence du mot « bœuf »

2.3 Calcul des adjectifs et noms dérivés du LVF

43 Une nouveauté de cette version du LVF par rapport aux précédentes est le fait que les dérivations adjectivales et nominales ainsi que le déverbal sont maintenant explicités plutôt que seulement signalés sous forme d'indicateurs comme on peut le constater par la Figure 1. Ces dérivés ont été obtenus par programme en utilisant les codes associés aux différents types de dérivés : adjectifs verbaux en *-ant, -é (-i, -u, ...)* et *-able* ainsi que les dérivés nominaux en *-age, -ment, -ion, -eur, -oir* et *-ure*.

44 Notre programme implante les indications données aux pages 14 à 19 de Dubois et Dubois-Charlier 1997. Les auteurs nous avaient confié qu'ils avaient développé un programme pour générer ces formes, ce qui est très vraisemblable étant donnés la finesse et les détails du codage. En page 17, les auteurs écrivent: « *Les irrégularités de formation (éteindre, éteignoir) ou les modifications orthographiques (dépecer, dépeçoir) sont traitées dans le programme de génération des formes* ». Ce programme n'ayant pu être retracé, nous en avons développé un nouveau, dont voici les grandes lignes.

45 Le principe du calcul des dérivations est simple, mais il doit tenir de plusieurs détails. En fonction de son groupe de conjugaison, un certain nombre de lettres sont enlevées à la forme infinitive du verbe pour obtenir son radical. Une terminaison est ajoutée au radical en fonction d'informations fournies dans les tables du document PDF. Il a fallu toutefois développer un certain nombre de *hacks* pour tenir compte de particularités

comme le changement d'un *c* en *ç* (*avançable* à partir d'*avancer*) ou *gu* en *g* (*largable* à partir de *larguer*). Certains dérivés, codés comme ayant une *formation irrégulière (endormir, endormissement ; gésir, gisement),* ne sont pas gérés par notre programme. Nous donnons maintenant plus de détails sur ce processus implanté avec un programme Python disponible auprès de Guy Lapalme.

46 On obtient le radical en enlevant la terminaison à l'aide du code de conjugaison (champ **gr**) : un chiffre et une lettre suivis de A, B ou Z qui indique l'auxiliaire. La page 19 donne des exemples de verbes dans chacun des 85 groupes (chiffre et lettre). Nous avons pu ainsi déterminer le nombre de lettres à supprimer de la forme infinitive du verbe pour retrouver le radical. Par exemple, pour le groupe **1a**, p. ex. *chanter*, on enlève 2 lettres pour obtenir *chant-,* alors que pour le groupe **5d**, p.ex. *craindre* ou *peindre*, on doit en enlever 3 pour obtenir *crain-* ou *pein-*.

47 En fonction des codes associés à chaque type de dérivés, on ajoute une terminaison et parfois aussi un préfixe. Par exemple, pour le verbe *défendre*, le code pour le dérivé adjectival en *-able* permet de calculer *défendable*, mais aussi *indéfendable*.

48 Les dérivés en *-ion* et *-eur* (p. 16) sont plus élaborés, car le codage indique qu'il faut encore enlever des lettres au radical avant d'y ajouter la terminaison. Par exemple, pour le verbe *indiquer* (avec radical *indiqu-*), le code **2A** demande d'enlever encore deux lettres et d'y ajouter *-cation* et *-cateur*, pour obtenir *indication* et *indicateur*. D'autres codes indiquent des dérivés multiples : par exemple, *composer* peut avoir deux dérivés en *-eur* : *composeur*, *compositeur* en plus des autres dérivés : les adjectifs *composable*, *composé* et *composant* et le nom *composition*.

49 Le programme de dérivation vérifie si les dérivés du LVF calculés par ce processus apparaissent dans le DEM. Si ce n'est pas le cas, ils sont préfixés par un astérisque. Dans la grande majorité des cas, ce sont des participes passés qui évidemment n'apparaissent pas comme entrée distincte dans le DEM, dans d'autres cas, cela peut indiquer un mot manquant dans le DEM. Cette vérification nous a facilité la mise au point de notre programme de dérivation, mais aussi permis de détecter une vingtaine d'erreurs mineures de codage que nous avons corrigées, ce qui est très peu sur plus de 25 000 entrées.

2.4 Un mode d'exploitation des données lexicales conforme aux intentions de M. Gross et de J. Dubois & F. Dubois-Charlier : associer les tables lexicales à des *grammaires locales* pour des tâches d'analyse syntactico-sémantique de corpus

50 Nous avons dit *supra* (Section 1.1) que les lexiques-grammaires du LADL ont été formatés pour le TAL et que le logiciel INTEX a été conçu à cet effet. Dans ce système, le *dictionnaire* est associé à des *grammaires locales*. Les deux types de ressources sont élaborés par l'utilisateur en fonction des ressources originales dont il dispose ainsi que de la finalité de ses recherches[11].

51 Pour illustrer le fait que ce type d'application reste très intéressant de nos jours, nous présentons ici une expérience d'annotation automatique syntaxique et sémantique d'un texte. Le dictionnaire que nous utilisons est une adaptation d'une portion de LVF. L'un de nous deux, Denis Le Pesant, travaille avec le logiciel de traitement de corpus NooJ[12], développé après la mort de Maurice Gross par Max Silberztein (*cf.* Silberztein

2010). Quelques autres expériences du même genre ont déjà été présentées par Sabatier et Le Pesant (2013), Le Pesant, Sabatier, Silberztein et Stéfanini (2014), Silberztein (2010) et Silberztein (2020 : dans ce numéro).

52 Les fonctionnalités du logiciel sont, entre autres : la désambiguïsation du sens des mots par prise en compte de leur distribution syntaxique, la reconnaissance des variantes de forme des mots (en particulier dans le cas des mots composés et des locutions) et des phrases, l'annotation morphologique, syntaxique et sémantique des corpus, et la fouille de textes.

53 Considérons la Classe « C2b » (Construction « T15a8 »)[13] de LVF. Ce sont verbes de communication dont la construction est « N0<qq> V N1<Objet> à N2<qq> » (ex. *quelqu'un* ordonne, *à quelqu'un, quelque chose* ou *de faire quelque chose,*). Appelons-les *verbes d'incitation.* Ce sont :

> commander 04, conseiller 01, déconseiller 01, défendre 05, demander 02, dicter 02, dire 03, édicter 02, enjoindre, imposer 01, interdire 01, interdire 06, intimer 02, offrir 03, ordonner 04, parler 11, permettre 01, prescrire 01, promettre 02, proposer 03, proposer 05, rappeler 08, réclamer 01, recommander 04, reconseiller, redéconseiller, redemander 02, redicter 02, réédicter, remontrer 02, represcrire 02, stipuler 02, redire 02, refuser 09, réordonner 02, répondre 02, repromettre 01, reproposer 02, souhaiter 02, interdire 05, refuser 08, persuader 03, suggérer 02, mander 02, interdire 04, permettre 01, permettre 02, préconiser, signifier 04.

54 La forme tabulaire de LVF (*cf. supra* Tableau 3) permet, par une opération de copier-coller et un petit ensemble d'opérations de traitement de texte, de construire un dictionnaire NooJ éventuellement modifié par rapport à l'original selon les choix de l'utilisateur. Voici un extrait du dictionnaire NooJ que nous avons utilisé pour cette démonstration. Nous aurions pu reprendre le formalisme original de LVF ; nous avons préféré y substituer un autre formalisme à nos yeux plus synthétique[14] :

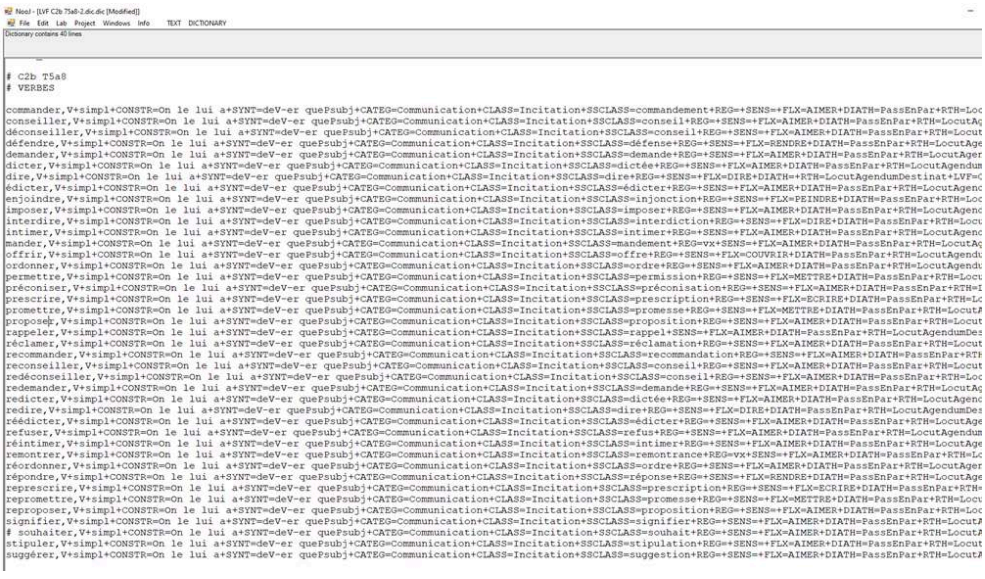


Figure 3 : Dictionnaire NooJ des verbes de la Classe C2b T5a8 de LVF

55 Ce dictionnaire, en même temps qu'un autre dictionnaire, un lemmatiseur, le *DM,* intégré à la version courante de NooJ qui reconnaît toutes les catégories lexicales et leurs éventuelles flexions[15], est ensuite appliqué à un corpus. A ce stade, les parties du discours et leur morphologie sont reconnues, mais dans un bruit considérable, à cause de la polysémie.

56 De fait, la majorité de verbes sont polysémiques (par exemple, selon LVF, le verbe *défendre* a cinq sens différents) et de nombreuses flexions sont elles-mêmes polysémiques. Pour déterminer quel est leur emploi dans un discours donné, il faut reconnaître et analyser leur contexte. Pour réduire les ambiguïtés, on construit une ou plusieurs *grammaires locales* qui vont définir des restrictions sur l'environnement des verbes, c'est-à-dire notamment sur la composition de leur structure argumentale. Comme ces graphes sont des transducteurs, les annotations supplémentaires éventuelles qu'ils produisent vont pouvoir compléter les annotations précédentes, issues du *DM*.

57 Supposons qu'on veuille relever dans un corpus ou un texte toutes les occurrences de *verbes d'incitation*, les analyser en tant que « verbes de *Communication*, Classe *Incitation* », et qu'on veuille aussi reconnaître leur contexte et l'analyser selon la fonction syntaxique et sémantique (par exemple « Sujet N, *Locuteur* », « Complément N, *Destinataire* », « Complément PROP à l'infinitif, *Objet* de l'incitation »). On construit alors une grammaire locale appropriée à ce vocabulaire, telle que :

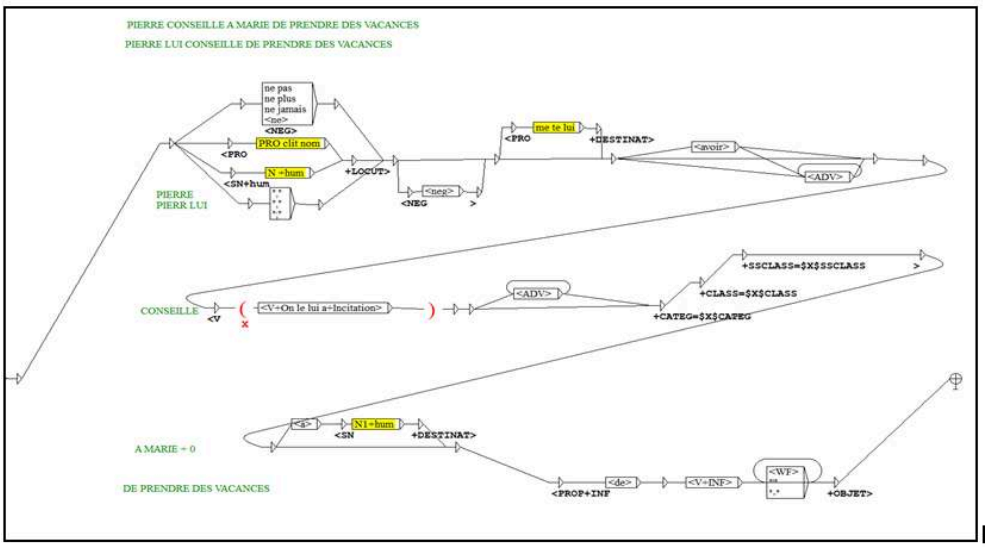


Figure 4 : Grammaire locale NooJ pour l'analyse et l'annotation des phrases à verbe appartenant à la Classe C2b T5a8 de LVF

58 Cette grammaire[16] est ensuite appliquée à un texte ou corpus. Nous choisissons pour cette démonstration le roman *Béatrix* (120.968 mots), extrait de *La Comédie humaine* de Balzac. L'analyse au moyen de NooJ, qui dure 56 secondes, produit la concordance de 32 occurrences que voici[17] :

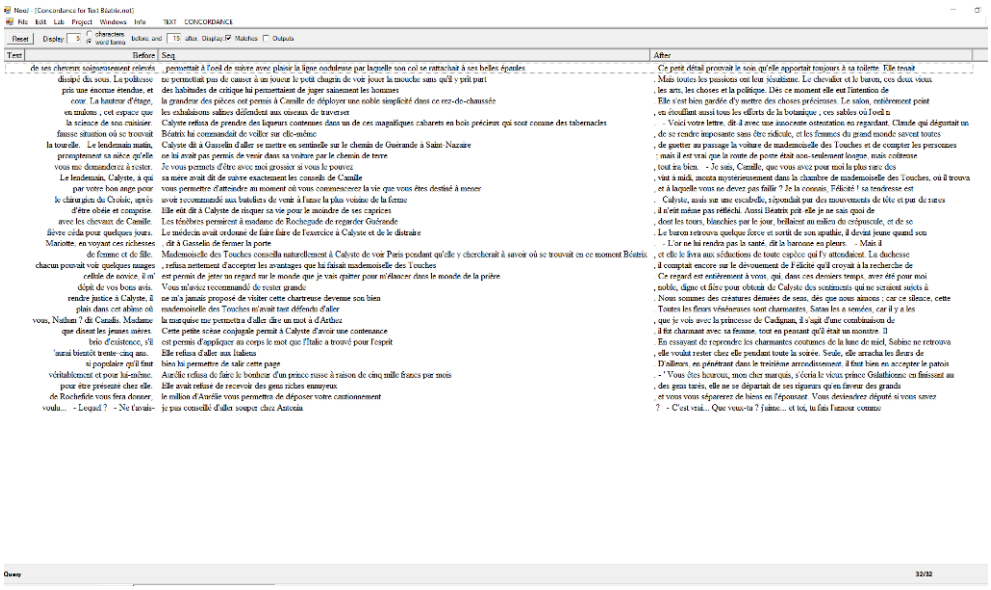


Figure 5 : Extrait de la concordance NooJ résultant de l'application de la grammaire locale de la Figure 3 à *Béatrix* de Balzac

59 Les annotations issues de l'application de la grammaire locale de la Figure 3 sont susceptibles d'être exportées dans le texte lui-même. Nous copions ci-dessous quelques extraits annotés en XML de *Béatrix* de Balzac :

> Le lendemain matin, <SN TYPE="hum" TYPE="LOCUT">Calyste</SN> <V CATEG="Communication " CLASS="Incitation " SSCLASS="dire">dit</V> à <SN TYPE="DESTINAT">Gasselin</SN> <PROP TYPE="INF" TYPE="OBJET">d'aller se mettre en sentinelle sur le chemin de Guérande à Saint-Nazaire</PROP>, de guetter au passage la voiture de mademoiselle des Touches et de compter les personnes qui s'y trouveraient.
> Toutes les fois que vous verrez un livre de musique ouvert sur le piano, vous me demanderez à rester. <PRO TYPE="LOCUT">Je</PRO> <PRO TYPE="DESTINAT">vous</PRO> <V CATEG="Communication " CLASS="Incitation " SSCLASS="permission">permets</V> <PROP TYPE="INF" TYPE="OBJET">d'être avec moi grossier si vous le pouvez</PROP>, tout ira bien.
> Elle était sûre d'être obéie et comprise. <PRO TYPE="LOCUT">Elle</PRO> eût <V CATEG="Communication " CLASS="Incitation " SSCLASS="dire">dit</V> à <SN TYPE="DESTINAT">Calyste</SN> <PROP TYPE="INF" TYPE="OBJET">de risquer sa vie pour le moindre de ses caprices</PROP>, il n'eût même pas réfléchi.
> Le docteur essaya de couper la fièvre avec du quinine, et la fièvre céda pour quelques jours. <SN TYPE="hum" TYPE="LOCUT">Le médecin</SN> avait <V CATEG="Communication " CLASS="Incitation " SSCLASS="ordre">ordonné</V> <PROP TYPE="INF" TYPE="OBJET">de faire faire de l'exercice à Calyste et de le distraire</PROP>. Le baron retrouva quelque force et sortit de son apathie, il devint jeune quand son fils se faisait vieux.

3. Le *Dictionnaire Electronique des Mots* : un projet de « lexique-grammaire total »

60 Le DEM est une ressource particulièrement méconnue. Jean Dubois et Françoise Dubois-Charlier ont travaillé sur ce projet jusque dans les derniers mois de leurs vies respectives, le laissant largement inaccompli. Nous le qualifions de « lexique-grammaire total » pour souligner qu'en intégrant le continent immense des mots non

prédicatifs, notamment les noms concrets, il constitue un complément essentiel aux lexiques-grammaires du LADL, qui ne prennent en compte que les mots prédicatifs.

3.1 Généralités sur le DEM

61 Les auteurs n'auront pas eu le temps de fournir une explicitation détaillée des codages. On trouvera cependant un certain nombre d'éléments d'information fournis par Françoise Dubois-Charlier au moment de la publication[18]. On pourra surtout lire leur dernier article (Dubois & Dubois-Charlier 2010), où ils décrivent leur projet avec l'exemple du domaine de la musique :

> Notre objet avec ce dictionnaire est double :
> – d'une part, essayer de faire en sorte que l'enchaînement des éléments contenus dans les rubriques pour un mot donné aboutisse à la formation d'une phrase élémentaire (axe syntagmatique) ;
> – d'autre part, essayer de faire en sorte que chaque étiquette employée dans une rubrique pour un mot se retrouve dans la même pour d'autres mots, définissant ainsi une classe de mots (axe paradigmatique).
> (Dubois & Dubois-Charlier 2010 : 179)

62 Dubois et Dubois-Charlier donnent un peu plus loin quelques explications sur leurs conventions de codage des propriétés à partir de l'exemple suivant :

Tableau 5

MOT	CONT	DOM	OP	SENS	OP1
accordéoniste	N q joue d	MUS	art	q joue d accordéon	C1c3

63 Les trois rubriques commentées ici sont CONT (pour « contexte » ou « contenu »), OP (pour « opérateur »), OP1 (pour un autre type d'opérateur). Le développement sur les termes de la musique illustrera ce qu'elles contiennent. Ainsi :

> *accordéoniste*, CONT = 'N qui joue de', OP = 'spéc', OP1 = 'C1c3' = « *un accordéoniste est une personne qui joue d'un instrument (défini dans la rubrique SENS), dont c'est la spécialité, ce qui en fait le sujet de verbes exprimant l'idée d'« émettre des sons à fonction expressive et esthétique* ».
> Il est important de noter que les étiquettes de la rubrique OP1 renvoient aux classes de verbes données dans LVF (1997), donc à un ensemble préétabli. Celles qui sont employées dans les rubriques CONT et OP, elles, ne renvoient pas à des listes préétablies, elles se dégagent peu à peu de l'analyse des termes à traiter en fonction de leur définition, de leur syntaxe, et à des fins de différenciation. Ainsi, de la considération des quelque 16.000 noms désignant des humains ont émergé, pour OP, des étiquettes comme 'spécialiste', 'habitant', 'amateur de', etc. De la considération des quelque 6 000 entrées adverbiales ont émergé, pour CONT, des étiquettes comme 'agir adv', 'parler adv', 'écrire adv', 'juger que', etc. (Dubois & Dubois-Charlier, 2010 :179)

64 Voici une image des premières entrées du DEM par ordre alphabétique des entrées. Elle illustre en particulier le parti pris d'intégrer systématiquement les expressions polylexicales au dictionnaire général :

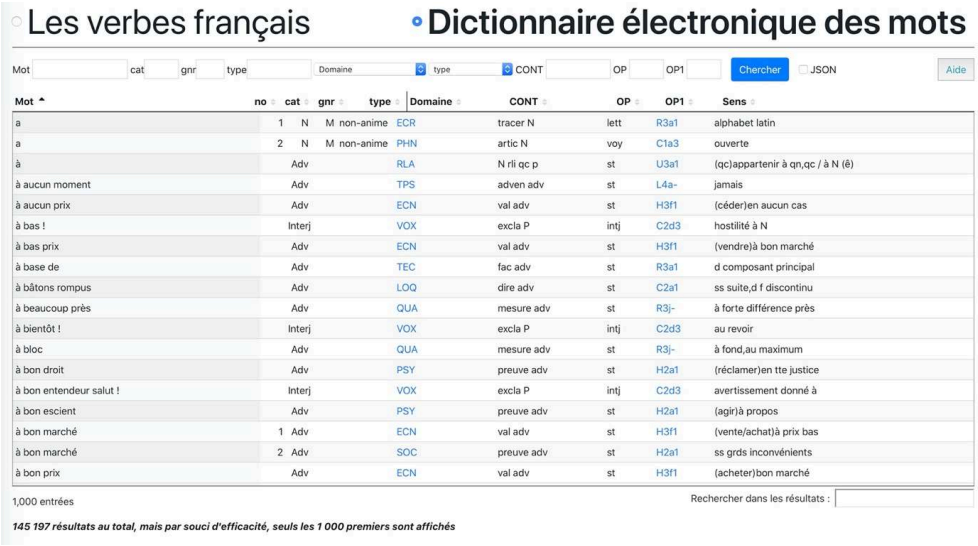


Figure 6 : les 18 premières des 145 197 entrées du DEM

65 On voit par ce court extrait que :

- toutes les parties du discours, sont représentées et explicitées à peu près en clair (dans le champ *cat*, « *N* » signifie « Nom » et « *Adv* » signifie « Adverbe » ; dans le même champ, on peut trouver par exemple *Vi* (« verbe intransitif », *Vp* (« verbe pronominal) ;
- le dictionnaire est loin de se limiter aux mots simples ; il n'y a que 3 mots simples sur 18 mots dans cet extrait[19] ;
- les différents emplois d'un même mot sont dégroupés ; par exemple, il y 2 emplois de l'expression *à bon marché* ;
- les registres de langue sont mentionnés : par exemple, pour l'entrée *a quia,* le « t » minuscule accolé au nom de domaine signifie : « registre littéraire » ;
- les mentions de domaines sont codées, mais une fois appris le codage, elles sont facilement mémorisables (RLA pour « relation », DRO pour « droit », TPS pour « temps », etc. ; il est à noter que VOX renvoie aux énoncés figés (appelés parfois pragmatèmes)[20] tels, dans la Figure 5, *à bas* qq ! et *à bientôt* !).

66 Le champ OP1 code des Classes de verbes de LVF. En mettant en relation les noms non prédicatifs du DEM avec les verbes prédicatifs de LVF les deux auteurs entendent, pour reprendre leur propre formule citée *supra* au début de cette section, « *essayer de faire en sorte que l'enchaînement des éléments contenus dans les rubriques pour un mot donné aboutisse à la formation d'une phrase élémentaire (axe syntagmatique)* ».

67 Revenons sur les expressions « *à bas x* ! » et « *à bientôt* ! » de la Figure 5. Elles sont reliées dans le champ OP1 à la Classe C2d3. On est renvoyé là à un certain nombre de verbes de « *dire en criant* » tels que *s'écrier* et *s'exclamer*. Autre exemple : le nom *a* (emploi 2), en tant que nom de phonème (*le phonème* « a »), est relié à la Classe C1a3, qui regroupe les verbes d'« émettre un type de paroles » tels que *aboyer, beugler*. Dans l'emploi 1, le nom « *a* », en tant que nom de graphème, est relié à la Classe R3a1 (du moins à une partie non négligeable de cette classe) qui sont des verbes transitifs directs de « réalisation d'un objet » tels que *coudre, dessiner, graver, façonner*.

68 Chaque ligne représente une structure de phrase. Par exemple, les entrées *a capella, a posteriori #2*, *a priori #*3 et *à aucun moment* sont à interpréter ainsi, respectivement :

chanter *a capella* c'est, dans le Domaine de la musique, chanter sans accompagnement
examiner *a posteriori* c'est, dans le Domaine des relations, examiner avec du recul

advenir *à aucun moment* c'est, dans le domaine du temps, n'arriver jamais

69 Ce n'est pas le lieu ici de décrypter les codages du DEM : ce serait une entreprise de longue haleine qui mettrait du reste en évidence un certain nombre de lacunes et d'incohérences caractéristiques d'une entreprise colossale laissée inachevée. D'ailleurs, ce ne serait pas nécessairement une tâche utile du point de vue syntactico-sémantique ou lexicographique. Nous développerons *infra* dans la Section 3.2 une autre piste d'utilisation du DEM, en tant que « matière première » exploitable non pas dans sa totalité, mais par sous-ensembles fragmentaires, selon les choix du chercheur. Mais donnons cependant au moins quelques illustrations du mode de codage des propriétés choisi par les auteurs du DEM[21].

70 Le principe d'interprétation des propriétés de chaque entrée est, on vient de le voir, de combiner les informations figurant dans les champs CONT, OP, DOM et SENS. Le codage des *verbes* est le suivant. Dans le champ CONT, le premier signe est « N majuscule », qui code l'existence d'un sujet. La suite code, en minuscule, un trait sémantique ; elle peut ou non comporter un symbole prépositionnel, notamment « p » (= *par*) et « v » (= *avec, au moyen de*). Exemples :

Tableau 6

M	CONT	DOM	OP	SENS
manger	N ali prod	ALI	prod	absorber ali
vaseliner	N soigne p	PHA	tech	enduire d vaseline

71 Ces entrées s'interprètent comme suit, par combinaison des informations contenues dans les différents champs : « *manger*, c'est se nourrir d'un *produit alimentaire*, c'est-à-dire *absorber un aliment* » ; et « *vaseliner*, c'est *soigner au moyen de* (*cf.* symbole « *p* ») une *technique pharmaceutique* qui consiste à *enduire de vaseline* ».

72 Dans le cas des *noms*, dans le champ CONT, le signe « N majuscule » ne figure pas en première position (comme c'est le cas dans les entrées verbales, *cf.* Tableau 6) et renvoie à l'entrée. Exemples :

Tableau 7

M	CONT	DOM	OP	SENS
yaourt	ali N	ALI	prod	lait caillé, yoghourt
assiette	ali v N	VAI	ins	pr manger aliments

73 Ces entrées s'interprètent comme suit, par combinaison des informations contenues dans les différents champs : « on mange du *yaourt*, *produit alimentaire* consistant en *lait caillé*, dit aussi *yoghourt* » ; et « on mange *avec* (*cf.* symbole « *v* ») une *assiette*, *instrument* de *vaisselle* fait *pour manger des aliments* ».

74 Les noms d'agent humain comportent eux aussi un N majuscule, mais en position initiale et suivi de la lettre « q » minuscule, comme le montre le premier des trois exemples suivants :

Tableau 8

M	CONT	DOM	OP	SENS
racketteur	N q commet	PEN	acte	q rackette, rançonne
racketter	N commet p	PEN	acte	rançonner
racket	commet N	PEN	acte	extorsion d fonds

75 L'interprétation de ces exemples est la suivante : un *racketteur* est quelqu'un qui commet un acte relevant du droit pénal consistant à racketter ou rançonner ; d'autre part, *racketter* c'est commettre un acte relevant du droit pénal par le fait de rançonner ; enfin, on commet un *racket* par un acte relevant du droit pénal et consistant à procéder à une extorsion de fonds.

76 Dans ces exemples, le lien morphologique entre les trois mots partageant le même radical s'effectue d'une part par la similitude des informations figurant dans les champs OP et DOM, d'autre part par des renvois possibles entre les entrées et les éléments de SENS. Quant au champ SENS, il renvoie à des entrées synonymiques (*rançonner, extorsion de fonds*), mais il est à noter que les renvois à d'autres entrées ne sont malheureusement pas formalisés.

3.2 Quelques propriétés remarquables du DEM

77 Les propriétés du DEM que nous allons évoquer ici sont, parmi d'autres, corrélées au caractère « totalisateur » du projet de Dubois & Dubois-Charlier. Le haut degré de valeur quantitative compense dans une certaine mesure les défauts qualitatifs de cette œuvre inachevée, que nous nous plaisons à qualifier de « matière première » de premier ordre pour l'étude du lexique du français.

3.2.1 Le DEM réserve le même traitement aux locutions qu'aux mots simples

78 Dans le DEM, les locutions ne sont pas, comme cela se pratique le plus souvent, répertoriées dans des dictionnaires séparés ou placées en annexe d'une entrée de mot simple. Elles sont mises sur le même plan que les mots simples : les deux catégories figurent dans la même table et peuvent, grâce au caractère informatisé de ce type de lexicographie, être triées par ordre alphabétique. Ce choix a le mérite de mettre en évidence le fait que, du point de vue syntactico-sémantique, rien ne distingue, à notre avis, les locutions des mots simples. D'autre part, il nous aide à mesurer jusqu'à quel point, dans de vastes régions du lexique, la proportion de locutions est importante. Par exemple, la proportion de locutions est très forte dans le domaine de la Psychologie. Considérons les 755 verbes psychologiques relevant du registre *standard.* Ils sont codés de la façon suivante dans le DEM :

Tableau 9

CONT	DOM	OP
N éprouver	PSY	sent

79 La proportion de locutions dans cette classe est de 79 %. Dans d'autres classes de verbes, la proportion de locutions est minoritaire, mais importante. Par exemple, il y a 237 verbes de Déplacement accessibles par la requête suivante :

Tableau 10

CONT	**DOM**	**OP**
N f mvt p	LOC	type

80 La proportion de locutions dans cette classe (y compris dans le registre familier) est de 38 %. Prenons un dernier exemple dans le domaine des verbes de Communication : on y compte 161 entrées ; 44 % de ces verbes sont des locutions. Ils sont accessibles par cette requête :

Tableau 11

CONT	**DOM**	**OP**
N dire pr	LOQ	dit

81 Il est à noter que LVF ne recense qu'une faible minorité des verbes locutionnels. A l'égard du vocabulaire des verbes français, le DEM compense donc une grave lacune de LVF.

82 Les inventaires de mots grammaticaux du DEM (prépositions, conjonctions, adverbes connecteurs), qui sont très majoritairement des locutions, sont particulièrement impressionnants par leur extension. Par exemple, dans le domaine des adverbes, prépositions et conjonctions de Temps, le DEM n'enregistre pas moins de 774 entrées, presque toutes locutionnelles. Prenons un autre exemple : le domaine de l'expression de la localisation compte 302 entrées prépositionnelles et adverbiales, très majoritairement locutionnelles. Elles sont définies par le codage suivant :

Tableau 12

CONT	**DOM**	**OP**
situer adv	LOC	st

83 Ces codes signifient à peu près ceci : « Dans le Domaine pragmatique du LIEU (*LOC*), ces lexèmes sont reliés, en position syntaxique adverbiale (*adv*), à des prédicats statifs (*st*)

de « situation » (*situer*) ». Voici par exemple un extrait de la liste des adverbes et prépositions dont le segment initial est « *à* » et « au, aux » :

> au bas de, au beau milieu de, au bord de, au bout du monde, au centre de, au cœur de, au coin du bois, au coin du feu, au confluent de, au coude à coude, au détour du chemin, au diable, au diable vauvert, au dos de, au droit de, au faîte de, au fond de, au frais, au grand air, au loin, au lointain, au milieu de, au niveau le plus élevé, au passage, au pied de, au sec, au sein de, au vert, au-dedans, au-delà, au-delà des mers, au-dessous, au-dessous de, au-dessus de, au-devant, au-devant de, aux alentours de, aux antipodes de, aux avant-postes, aux confins de, aux côtés de, aux environs de, aux pieds de, aux portes de, aux premières loges, aux quatre coins de, aux quatre vents, etc.

84 Cette présentation pourrait donner à penser que les catégories ne sont guère subdivisées. Ce n'est pas toujours le cas, grâce aux notations de para-synonymie du Champ *SENS* qui permettent de constituer des sous-classes sémantiques. Illustrons ce fait par un extrait du DEM :

Tableau 13 : Extraits de la liste des prépositions de localisation du DEM

MOT	CONT	DOM	OP	SENS	OP1
au dos de	situer adv	LOC	st	au verso de	L3a1
au milieu de	situer adv	LOC	st	au centre de loc	L3a1
aux environs de	situer adv	LOC	st	aux alentours de loc	L3a1
aux confins de	situer adv	LOC	st	aux limites de	L3a1
à ciel ouvert	situer adv	LOC	st	(ê) à l'air libre	L3a1
à fleur d'eau	situer adv	LOC	st	(ê) à la surface de l'eau	L3a1
à égale distance	situer adv	LOC	st	(ê) à même intervalle	L3a1
à découvert	situer adv	LOC	st	(ê) à nu, non protégé	L3a1

85 En dépit de ces exemples, force est d'admettre que la granularité sémantique, dans le DEM, n'est guère fine dans un grand nombre de domaines. Le DEM est selon nous avant tout une « matière première » de premier ordre disponible pour des recherches ultérieures plus spécifiées (Cf. *infra* Section 3.2.3).

3.2.2 Le DEM rend compte de tous les registres et incorpore les termes de spécialité

86 Dans le DEM, les registres de langue et les variations régionales du français sont codées dans le champ DOM (domaine) de la façon suivante :

> Niveaux de langue : f (familier), p (populaire), v (vieux), t (littéraire)
> Régionalismes : b (belge), c (canadien, québécois), s (suisse)

87 Les verbes psychologiques relevant du registre standard, hors registres *argot* et *vulgaire*, sont codés de la façon suivante dans le DEM :

Tableau 14

CONT	DOM	OP
N éprouver	PSY	sent

88 On constate que ce sont des locutions dans une proportion de 92%.

89 Jean Dubois fut un des coordinateurs des équipes qui ont élaboré le *Grand Dictionnaire Encyclopédique Larousse*. C'est cela qui l'a probablement encouragé à donner au DEM un caractère en partie terminologique. Ainsi le DEM incorpore-t-il tous les vocabulaires de spécialité. Par exemple, dans le Champ Domaine, la mention REP rassemble 578 noms de reptiles, dont 81 noms de reptiles fossiles ! Autre exemple, le Domaine CHM (chimie), le DEM rassemble 5147 termes. Ce parti pris pourrait surprendre, d'autant que le DEM n'est évidemment pas une ressource terminologique fiable, ne serait-ce qu'en raison du fait qu'il n'est pas et ne sera jamais actualisé. Il témoigne tout au moins du caractère « totalisateur » du projet de Dubois et Dubois-Charlier.

3.2.3 La mise en relation des noms non prédicatifs avec les prédicats appropriés dont ils sont les arguments

90 Dans Gross (1990), Maurice Gross évoque le traitement des noms dans le lexique-grammaire. Les *noms prédicatifs* ont à cette époque déjà fait l'objet de nombreux travaux, et Gross fait la liste d'ouvrages de ses collaborateurs sur divers types de constructions à verbes supports *faire, avoir, donner-recevoir, être en, avoir-prendre-perdre* (*cf.* entre autres J. Giry-Schneider 1978, 1987, G. Gross 1989, R. Vivès 1983). « *Tous ces travaux*, écrit Gross, *démontrent la possibilité de représenter ces mots dans un lexique-grammaire* ». Et d'ajouter :

> Seuls semblent échapper à cette description les substantifs **concrets** (e.g. *chaise, notaire, homme des bois, pomme de terre*), qui ne donnent pas lieu à des phrases élémentaires syntaxiquement significatives, et présentent d'autres problèmes de représentation (souligné par nous). (Gross, M. 1990 : 49)

91 Les noms non prédicatifs, c'est-à-dire les noms qui n'ont pas de structure argumentale, sont donc absents des lexiques-grammaires du LADL. Il en va tout autrement dans le cas du DEM, qui intègre la totalité de la langue, y compris donc des dizaines de milliers de noms concrets. Une question se pose donc, du point de vue théorique : le DEM, en intégrant la multitude des lexèmes non prédicatifs s'écarte-t-il du projet des *lexiques-grammaires* issu des grammaires de Harris ?

92 Les travaux de Gaston Gross et de son équipe le LLI[22], complètement indépendants de ceux de Dubois et Dubois-Charlier, montrent qu'un traitement des mots non prédicatifs est possible dans le cadre théorique des grammaires de Harris. La notion théorique qui permet cette approche est celle de *sélection lexicale*, qui découle de celle de *distribution du prédicat* : étant donné un prédicat quelconque, il y a des restrictions sur le choix des éléments qui constituent sa structure argumentale. A partir des classes de prédicats, on peut définir les classes de mots (prédicatifs ou non prédicatifs) qu'ils sélectionnent (ce fut là la tâche des chercheurs du LADL). Mais le chemin inverse est possible : définir les

classes d'arguments en prenant en compte les prédicats qui les sélectionnent spécifiquement.

93 Ces classes d'arguments furent appelées par Gaston Gross *classes d'objets* ; les prédicats associés aux *classes d'objets* furent appelés *prédicats appropriés*. Les classes d'objets peuvent être des classes de prédicats ; par exemple les classes de noms de *spectacles* (ex. *représentation* Adj, *séance de* N) sont sélectionnées par des prédicats appropriés tels que *donner, assister* à. Mais les classes d'objets peuvent être également des classes de noms non prédicatifs, et c'est à ce cas particulier que nous sommes en train de nous attacher : par exemple, les classes d'objets de *nourriture et boissons* ont pour prédicats appropriés des verbes tels que *prendre* (dans un de ses multiples emplois), *goûter* et *déguster*[23].

94 Revenons au DEM. Il se trouve que tous les noms, qu'ils soient ou non prédicatifs sont corrélés, grâce au Champ OP1, à des Classes de LVF (*cf.* Section 1.3 et Figure 5). Il s'ensuit *de facto* que les noms non prédicatifs sont corrélés avec leurs *prédicats appropriés*. Cette constatation doit être assortie d'une réserve : la corrélation est souvent approximative, car les verbes de chaque Classe de LVF mentionnée dans le champ OP1 ne correspondent pas tous à leur corrélat du DEM.

95 L'un de nous deux, Denis Le Pesant, est frappé, plus de 25 ans après la publication d'un article d'illustration de la théorie des *classes d'objets* (Le Pesant 1994) qui faisait l'esquisse d'une sorte de thésaurus de l'*Ecriture* et des *Supports de l'écriture*, par l'élévation considérable du niveau de faisabilité du projet que permettrait aujourd'hui l'exploitation conjointe de LVF et du DEM.

96 Considérons de façon privilégiée, comme l'avait fait Le Pesant (1994), les prédicats appropriés des noms de texte (ex. *lire*) par rapport aux classes d'objets d'*écrits*. Le verbe *lire* est le prédicat approprié le plus général et il n'avait été jugé nécessaire de chercher d'autres prédicats appropriés des noms d'écrits que *lire, relire, déchiffrer* et *décrypter*. Si la base de données de LVF avait été disponible, il aurait suffi de prélever dans la Classe P3b les verbes et noms déverbaux qui nous intéressent, ceux dont l'Opérateur est « *scrut texte* », à savoir : *analyser* (*analyse*), *approfondir* (*approfondissement*), *compulser, creuser, décortiquer, dépiauter, disséquer, éplucher, explorer, interroger, lire* (*lecture*), *relire* #1, *relire* #2 (pour corriger), *relecture,* etc.

97 Pour ce qui est des prédicats d'*écriture* il aurait été facile de recenser, en se référant à la Classe R4a de LVF, de nombreuses modalisations de *écrire* et *rédiger*, telles que *accoucher de, bâtir, composer* (*composition*), *commettre, compiler* (*compilation*), *composer* (*composition*), *esquisser* (*esquisse*), *expédier, pisser, pondre, trousser* etc. On aurait aussi trouvé dans LVF les prédicats de publication (*éditer, publier ; édition, publication*, etc.).

98 Passons aux classes d'objets sélectionnées par les classes de prédicats de *lecture* et d'*écriture*. On aurait pu faire d'abord dans le DEM la requête suivante :

Tableau 15

CONT	DOM	OP
lire N		txt

99 On obtient alors la liste des noms référant à ce qui se lit sans être pour autant nécessairement un texte suivi. On les trouve notamment les Domaines suivants :

ADMinistration (Ex. *feuille de soins*), ASTronomie (Ex. *éphéméride*), BIOlogie (Ex. *spermogramme*), DROit (Ex. *charte*), ECN (Ex. *registre*), ENSeignement (Ex. *encyclopédie*), LITtérature (Ex. *roman*), MEDecine (Ex. électro-cardiogramme), MUSique (Ex. *livret d'opéra*), PREsse (Ex. magazine), RELIGion (Ex. *missel*).

100 Passons maintenant aux classes d'objets d'*Ecrits*. Il s'agit de noms non prédicatifs. Si on avait eu à disposition le DEM, en effectuant la requête suivante :

Tableau 16

CONT	DOM	OP
écrire N		txt

101 on pouvait relever près de 1000 noms de *textes rédigés*, et ensuite affiner le tri en fonction des Domaines suivants :

> ADMinistration (Ex. *procès-verbal*), CINéma (Ex. *scénario*), DROit (Ex. *testament*), ECNomie (Ex. *devis*), ENSeignement (Ex. *dissertation*), LINguistique (Ex. *thésaurus*), LITtérature (Ex. *épopée, pamphlet*), METRique (Ex. *tercet*), POLitique (Ex. *tract*), PREsse (Ex. *éditorial*), RLGion (Ex. *encyclique*).

102 L'article de Le Pesant (1994) traitait aussi des classes d'objets de *supports de l'écriture*. Elles sont accessibles dans le DEM avec la recherche suivante :

Tableau 17

CONT	DOM	OP
écrit sur N	x	txt

103 Les principaux Domaines concernés sont : ADMinistration (*carte-grise*), ECN (*bulletin-réponse*), RELIGion (*phylactère*) et ECRiture (*calepin*).

104 D'autres parties du DEM rassemblent les noms d'auteurs (ceux qui écrivent, tels *auteur* et *journaliste*) et les noms de signes graphiques (qui se « lisent » également, mais dans un autre emploi du verbe *lire*).

105 Force est de constater que les recherches de Gaston Gross et de Jean Dubois & Françoise Dubois-Charlier, pour indépendantes les unes des autres qu'elles soient, ont un point commun qui les rattache étroitement aux principes théoriques de Harris. Il s'agit de l'intention déjà citée de Dubois & Dubois-Charlier (*cf. supra* Section 3.1) et que nous répétons ici :

> – d'une part, essayer de faire en sorte que l'enchaînement des éléments contenus dans les rubriques pour un mot donné aboutisse à la formation d'une phrase élémentaire (axe syntagmatique) ;
> – d'autre part, essayer de faire en sorte que chaque étiquette employée dans une rubrique pour un mot se retrouve dans la même pour d'autres mots, définissant ainsi une classe de mots (axe paradigmatique).
> (Dubois & Dubois-Charlier 2010 : 179)

Conclusion

106 Nous avons, dans cet article sur les deux dernières productions de Jean Dubois et Françoise Dubois-Charlier, visé six objectifs :

- rappeler l'inscription de ces dictionnaires dans le courant théorique initié par Zellig Harris et continué par Maurice Gross et son équipe, le LADL : celui des *grammaires distributionnelles et transformationnelles* ;
- constater que ce courant théorique se caractérise par la pratique d'un nouveau type de lexicographie : une lexicographie informatisée disponible pour le TAL ;
- effectuer un rapprochement entre l'entreprise du DEM et les travaux de Gaston Gross et de son équipe, le LLI ;
- montrer que l'ouvrage colossal mais inachevé qu'est le DEM constitue, en dépit de ses obscurités et imperfections, une sorte de « matière première » d'une grande valeur pour les recherches linguistiques et pour le TAL ;
- mettre en évidence la complémentarité des *Verbes Français* et du *Dictionnaire Electronique des Mots*
- renouveler et améliorer les interfaces de consultation de LVF et du DEM mises à la disposition du public par l'Université de Montréal.

107 Sur ce dernier point, la publication de cet article aura été l'occasion d'une refonte en profondeur des outils de consultation disponibles sur un site de l'Université de Montréal (*cf. supra,* Introduction). Il est apparu indispensable que soit développée une approche intégrée de LVF et du DEM, que nous avions jusqu'à présent considérés comme indépendants alors qu'ils doivent être considérés comme complémentaires l'un de l'autre, ce qui était d'ailleurs incontestablement le cas dans l'esprit des deux auteurs. Nous espérons que ces nouveaux outils de consultation contribueront à l'augmentation de la notoriété des dictionnaires informatisés du français de Jean Dubois et Françoise Dubois-Charlier auprès de la communauté des chercheurs, non seulement en linguistique mais aussi en TAL.

BIBLIOGRAPHIE

Bédaride, P. (2012). « Raffinement du Lexique des Verbes Français », in *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 2 : TALN. ATALA/AFCP : 155-168. [http://www.aclweb.org/anthology/F/F12/F12-2012]

Boons, J.-P., Guillet, A., Leclère, Ch. (1976). *La structure des phrases simples en français*, *I Constructions intransitives*. Genève : Droz.

Chevalier, J.-C., Encrevé, P. (2006). *Combats pour la linguistique, de Martinet à Kristeva*. Lyon : ENS Editions.

Daladier, A. (1990) éd. *Les grammaires de Harris et leurs questions*, in *Langages* n° 99. Paris : Larousse.

Danlos, L., Nakamura, T. Pradet, Q. (2014). « Vers la création d'un Verbnet français », in TALN-RECITAL 2014, *Workshop FondamenTAL 2014*, « Ressources lexicales et TAL. Vue d'ensemble sur les dictionnaires électroniques de Jean Dubois et Françoise Dubois-Charlier »

Devlin, J., Chang M.-W., Lee, K., Toutanova, K. (2019). « BERT: Pre-training of deep bidirectional transformers for language understanding », in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, pp. 4171-4186. Minneapolis, Minnesota, June 2019.

Dostie, G., Tutin, A. (dir) 2019. *Les phrases préfabriquées*, in *Cahiers de Lexicologie* n° 114. Paris : Classiques Garnier.

Dubois, J. et Dubois-Charlier, F. (1997 a). *Les verbes français* (LVF). Version informatisée (environ 26 000 entrées). Version LVF +1 (Micosoft-Excel) : https://www.modyco.fr/fr/base-documentaire/ressources/jean-dubois.html. Versions JSON et XML : http://rali.iro.umontreal.ca/rali/?q=fr/versions-informatisees-lvf-dem

Dubois, J. et Dubois-Charlier, F. (1997 b). *Les Verbes français* (LVF), 428 pages. Paris : Larousse.

Dubois, J. et Dubois-Charlier, F. (2011). « La combinatoire lexico-sémantique dans le Dictionnaire Electronique des Mots. Les termes du domaine de la musique à titre d'illustration », in Leeman et Sabatier (éds), *Langages* n° 179-180. Paris : Paris : Armand Colin. 31-56.

Dubois, J. et Dubois-Charlier, F. (2020). *Dictionnaire Electronique des Mots* (DEM). Version Micosoft-Excel : https://www.modyco.fr/fr/base-documentaire/ressources/jean-dubois.html. Versions JSON et XML : http://rali.iro.umontreal.ca/rali/?q=fr/versions-informatisees-lvf-dem

Dutoit, D., François, J. (2007). « *Changer* et ses synonymes majeurs entre syntaxe et sémantique », in *Langue Française* n° 153, pp. 40-57. Paris : Larousse.

François, J. (2017). « A quel titre le passif concerne-t-il les lexicographes ? *Les Verbes Français* et la perspective passive », in *Etudes de Linguistique Appliquée* n° 187, pp. 297-309. Paris : Klincksieck.

François, J., Le Pesant, D., Leeman, D. (2007). « Présentation de la Classification des Verbes Français, de Jean Dubois et Françoise Dubois-Charlier », in *Langue Française* 153, pp. 3-19. Paris : Larousse.

Giry-Schneider, J. (1987). *Les prédicats nominaux en français. Les phrases simples à verbe support*. Genève-Paris : Droz.

Gross G. (1989). *Les constructions converses du français*. Genève-Paris : Droz.

Gross G. (1992). « Forme d'un dictionnaire électronique », in A. Clas et H. Safar (éd.), *L'environnement traductionnel*, pp. 255-271. Sillery : Presses de l'Université de Québec.

Gross G. (1994). « Classes d'objets et description des verbes », in *Langages* n° 115, pp 15-30. Paris : Larousse.

Gross, G. (1998). « Pour une véritable fonction synonymie dans un traitement de texte », in *Langages* n° 131, pp. 103-114. Paris : Larousse

Gross, M. (1975). *Méthodes en syntaxe, régime des constructions complétives.* Paris : Hermann.

Gross. M. (1981). *Les bases empiriques de la notion de prédicat sémantique*, in *Langages* n° 63, pp. 7-52. Paris : Larousse.

Guillaume, B., Fort, K., Perrier, G., Bédaride, P. (2014). « Mapping the Lexique des Verbes du Français (Lexicon of French Verbs) to a NLP Lexicon using Examples », in *International Conference on Language Resources and Evaluation* (LREC), May 2014. Reykjavik, Iceland.

Guillet, A., Leclère, Ch. (1992). *La structure des phrases simples en français, II Constructions locatives*. Genève : Droz.

Harris, Z. (1976). *Notes du cours de syntaxe*. Paris : Le Seuil.

Johnson, M., Schuster M., Le, Q. V., Krikun M., Wu Y., Chen Z., Thorat N., Viegas, F., Wattenberg M., Corrado, G., Hughes M., Dean, J. (2017). « Google's multilingual neural machine translation system: Enabling zero-shot translation ». *Transactions of the Association for Computational Linguistics* n° 5:339–351.

Leeman, D., Sabatier, P. (2010) (éds). *Empirie, Théorie, Exploitation : le travail de Jean Dubois sur les verbes français*, in *Langages* n° 179-180. Paris : Larousse.

Le Pesant D. (1994). « Les compléments nominaux du verbe *lire* », in *Langages* n° 115, pp. 31-46. Paris : Larousse.

Le Pesant D., Mathieu-Colas M. (1998). « Introduction aux classes d'objets », in *Langages* n° 131, pp. 6-33. Paris : Larousse.

Lewis , N. (1978). *New Rogets Thesaurus*. Berkley Publishing Group.

Mathieu-Colas, M. (1998). « Illustration d'une classe d'objets : les voies de communication », in *Langages* n° 131, pp. 77-90. Paris : Larousse.

Mazziotta, N. (2014). « Analyse exploratoire des propriétés sémantiques et syntaxiques des verbes « psychologiques » (classe P) dans *Les Verbes Français* », in *JADT 2014 : 12èmes Journées internationales d'Analyse statistique des Données Textuelles*. Paris : Université Paris 3.

Mel'čuk, Igor. (1997) *Vers une linguistique Sens-Texte*. Leçon inaugurale (faite le Vendredi 10 janvier 1997), Collège de France, Chaire internationale, 43 pages. http://olst.ling.umontreal.ca/pdf/melcukColldeFr.pdf

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013). « Distributed representations of words and phrases and their compositionality », in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 3111-3119.

Miller, G. A. (1995). « WordNet: A Lexical Database for English », Communications of the ACM Vol. 38, N° 11: 39-41.

Norman Lewis (1978). *The New Roget's Thesaurus of the English Language in Dictionary Form*. New York : G.P. Putnam's Sons.

Paroubek, P., Chaudiron, S, Hirschman, L. (2007). « Principles of Evaluation in Natural Language Processing », *Traitement Automatique des Langues*, ATALA, 2007, 48 (1), pp.7-31.

Perrin, D. (1989). « Automates *et al*gorithmes sur les mots », in *Annales des Télécommunications,* Tome 44, n° 1-2, pp. 20-33.

Robert, P. (1980). *Dictionnaire alphabétique et analogique de la langue française*. Paris : Société du Nouveau Littré.

Ruppenhofer J., Ellsworth M., Petruck M.R.L., Johnson C.R., Baker C. F., Scheffczyk J.: *FrameNet II: Extended Theory and Practice* (2016.) https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf

Sabatier, P., Le Pesant, D. (2013). « Chapitre 5. Les dictionnaires électroniques de Jean Dubois & Françoise Dubois-Charlier et leur exploitation en TAL », in Gala, N. et Zock, M. *Ressources Lexicales*, pp. 153-186. Amsterdam : John Benjamins.

Salkoff, M. (1980). *Analyse syntaxique du français : Grammaire en chaîne*. Amsterdam : John Benjamins.

Silberztein, M. (2010). « La formalisation du dictionnaire *LVF* avec NooJ et ses applications pour l'analyse automatique de corpus », in *Langages* n° 179-180, pp. 31-56. Paris : Armand Colin.

Silberztein, M. (2015). *La formalisation des langues. L'approche de NooJ*. London : Iste Editions.

Trouilleux, F. (2018). « Le DM, a French Dictionary for NooJ », in Vučković, Bekavac and Silberztein, *Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference*. Cambridge, Scholars Publishing, pp.16-28, 2012, 1-4438-3711-3.ff hal-00702348f.

Van den Eynde, K., Mertens, P. (2006). *Dictionnaire de valence des verbes français : manuel d'utilisation* <http://bach.arts.kuleuven.be/dicovalence/manuel_100625.pdf>

Vivès, R. (1983). *Avoir, prendre perdre : constructions à verbe support et extensions aspectuelles*. Thèse de doctorat, Université Paris VIII, LADL.

Wang, D., Wang X., Lv, S. (2019). « An overview of end-to-end automatic speech recognition », in *Symmetry*, 11(8).

NOTES

1. Le LADL (*Laboratoire d'Automatique Documentaire et Linguistique*) est l'unité de recherche du CNRS et de l'Université Paris-Jussieu que Maurice Gross dirigea du début des années 1970 jusque peu de temps avant sa mort.

2. Tout prédestinait Maurice Gross, à l'époque des débuts de l'informatique, à donner à ses dictionnaires un format approprié au traitement informatique. En effet, avant d'être linguiste, il fut mathématicien et ingénieur informaticien. Très lié à Marcel-Paul Schützenberger, un des principaux fondateurs de l'informatique théorique et coauteur du théorème dit « de Chomsky-Schützenberger », il fut recommandé par Chomsky à Z. Harris, qui à son tour souhaitait donner une forme mathématique correcte à ses grammaires. A ce sujet on pourra consulter Chevalier & Encrevé (2006 : 251-261).

3. Pour une meilleure lisibilité, les champs consacrés aux variantes dérivationnelles (ex. suffixations en *–ant* et en *–able*) ont été supprimés dans cette image.

4. Verbes simples (67 tables regroupant 13 872 entrées) ; noms prédicatifs simples et composés (81 tables regroupant 14 271 entrées) ; expressions figées principalement verbales et adjectivales (69 tables regroupant 39 628 entrées) ; adverbes simples et figés (32 tables regroupant 10 448 entrées)

5. Il est téléchargeable en format Microsoft-Excel à partir de : http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Telechargement.html

6. Dans sa *Syntaxe de l'adverbe* (1987), M. Gross présente un lexique-grammaire de 6400 adverbes. Ils sont répartis entre 16 classes : une seule classe de 520 adverbes simples, tel *soudain,* et 15 classes de 5880 figés, tel *en désespoir de cause*.

7. Pour une meilleure lisibilité, les champs concernant les registres et les régionalismes ont été éliminés de cette image, ainsi que les champs flexionnels et dérivationnels.

8. JSON est l'acronyme de JavaScript Object Notation : https://www.json.org/json-fr.html

9. http://jsonlines.org

10. Cette interface JSON ainsi qu'une documentation détaillée sont accessibles à : http://rali.iro.umontreal.ca/rali/?q=fr/versions-informatisees-lvf-dem

11. Les requêtes les plus simples peuvent être traitées par de simples expressions régulières plutôt que par des grammaires locales.

12. NooJ est téléchargeable à partir de http://www.nooj-association.org/. En corrélation avec le présent article, un petit dictionnaire NooJ des verbes de la Classe C2b ainsi qu'une grammaire NooJ sont disponibles à http://denis.lepesant.pagesperso-orange.fr/ de façon à permettre la reproduction de la démonstration qui va suivre.

13. Le symbole *T15a8* code les propriétés syntaxiques suivante : verbe *transitif direct* (« T »), à *sujet humain* (« 1 »), à proposition subordonnée complétive *que* P ou infinitive (« 5 »), à complément second indirect en *à* (« a »), relié morphologiquement à un nom (« 8 »).

14. Par exemple, dans le champ CONSTR le codage « On le lui a » amalgame les informations suivantes : sujet humain, complément premier non humain, complément second humain datif.

15. Le *DM* est dû à François Trouilleux (cf. Trouilleux 2018).

16. Les éléments de couleur jaune de la grammaire renvoient à des sous-grammaires.

17. Sur ces 32 occurrences, 9 correspondent à une analyse erronée, à chaque fois à cause du verbe *permettre*, qui peut avoir, avec un autre sens, un sujet non humain, comme dans « *cet espace que les exhalaisons salines défendent aux oiseaux de traverser* ». Ce fait met en évidence un défaut dans la grammaire locale de la Figure 4.

18. Ce document est disponible à http://rali.iro.umontreal.ca/rali/?q=fr/versions-informatisees-lvf-dem.

19. Dans le « Grand Robert » (Robert 1980), les premières entrées concernent les différents emplois du nom *a,* du préfixe *a-,* et de la préposition *à* ; ensuite figure la locution *ab absurdo* (en 7 ème position) et le nom *abaca* (en 8ème position par ordre alphabétique). Dans le DEM, *ab absurdo* est absent (au profit de *ab intestat, ab irato* et *ab ovo*, ce dernier mot figurant en 956ème position), et le nom *abaca* figure en 957ème position par ordre alphabétique.

20. Cf. les articles d'Agnès Tutin et de Denis Le Pesant sur les *Phrases Préfabriquées des Interactions*, dans Dostie & Tutin (2019).

21. Le caractère abscons des codages s'explique en partie par le fait que les auteurs ont toujours travaillé sur une base de données ancienne (Dbase) qui exigeait un nombre limité de caractères par champ.

22. Laboratoire de Linguistique Informatique (LLI), CNRS & Université Paris 13. Fondateur et directeur : Gaston Gross.

23. Plusieurs publications illustrent la notion de *classe d'objets* en esquissant la forme d'un dictionnaire articulant des classes de noms non prédicatifs avec les classes correspondantes de prédicats appropriés. C'est le cas notamment des articles de Gaston Gross (1992 et 1994), Le Pesant (1994), Le Pesant & Mathieu-Colas (1998) et Mathieu-Colas (1998).

RÉSUMÉS

Notre article décrit la structure des ressources lexicales *Les Verbes Français* (LVF) et le *Dictionnaire Électronique des Mots* (DEM) élaborées pendant plusieurs années par Jean Dubois et Françoise Dubois-Charlier. Nous suggérons ensuite des utilisations possibles de ces ressources pour le traitement automatique de la langue (TAL). Compte-tenu du fait que LVF a déjà fait l'objet de plusieurs travaux au cours des dernières décennies, nous insistons sur le DEM, une ressource linguistique particulièrement mal connue qui peut être considérée comme la synthèse des travaux lexicographiques de Dubois et Dubois-Charlier. Le DEM souffre d'être resté inachevé, mais son extension peu commune (près de 150 000 entrées) et surtout ses corrélations avec LVF

en font une source de données lexicales de premier ordre pour la linguistique du français et pour le TAL. Nous présentons de nouvelles versions du LVF et du DEM au format JSON avec une nouvelle interface de consultation de ces dictionnaires. Nous espérons de la sorte favoriser la diffusion de ces ressources lexicales auprès de la communauté des chercheurs en lexicologie, en lexicographie et en TAL.

This paper presents the structure of two French lexical resources, *Les Verbes Français* (LVF) and *Dictionnaire Électronique des Mots* (DEM), created over many years by Jean Dubois and Françoise Dubois-Charlier. Applications of these resources for Natural Language Processing (NLP) are then sketched. Given the fact that LVF has already been studied quite extensively over the last decades, DEM is described in details as it can be considered as the synthesis of the lexicographic works of Dubois and Dubois-Charlier. DEM was unfortunately left as a work in progress, but its extension (almost 150 000 entries) and especially its links with LVF constitute an important French lexical resource for NLP applications. A new JSON based format of LVF and DEM is also presented with an integrated query interface. This should promote the use of these resources by researchers in lexicology, lexicography and NLP.

INDEX

Mots-clés : lexicologie, lexicographie, TAL (Traitement Automatique des Langues), grammaires distributionnelles et transformationnelles, Jean Dubois, Françoise Dubois-Charlier, interfaces de consultation de données lexicales, XML, JSON
Keywords : lexicology, lexicography, NLP (Natural Language Processing), distributional and transformational grammars, Jean Dubois, Françoise Dubois-Charlier, lexical database query interface, XML, JSON.

AUTEURS

GUY LAPALME

Université de Montréal

DENIS LE PESANT

Université de Paris Nanterre