Le dictionnaire DEM dans NooJ

Max Silberztein
ELLIADD, Université de Franche-Comté, 30 rue Mégevand, 25000 Besançon
max.silberztein@univ-fcomte.fr

Résumé. Nous avons intégré le *Dictionnaire Electronique des Mots* de Jean Dubois et Françoise Dubois-Charlier dans la plateforme linguistique NooJ. Nous montrons l'intérêt de ce dictionnaire pour les applications du TAL.

Abstract. We have integrated Jean Dubois et Françoise Dubois-Charlier's *Dictionnaire Electronique des Mots* in the NooJ linguistic software. We discuss the applications for Natural Language Processing applications.

Mots-clés : Dictionnaire électronique. NooJ.

Keywords: Electronic Dictionaries. NooJ.

1 Introduction

Le travail décrit ici en hommage à notre collègue et ami Paul Sabatier a pour double but de décrire avec exhaustivité et une précision absolue (i.e. de *formaliser*) le vocabulaire du français, et de construire des applications du TAL pour ces ressources linguistiques. Pour décrire le vocabulaire du français, nous avons implémenté avec la plateforme linguistique NooJ[1] le dictionnaire *Les Verbes Français* (**LVF**) et le *Dictionnaire Electronique des Mots* (**DEM**) construits par Jean Dubois et Françoise Dubois-Charlier, et récemment publiés[2].

2 Le dictionnaire LVF

Le dictionnaire *Les Verbes Français* (LVF) est disponible depuis 2010[3] et a été adapté pour être utilisé par la plateforme NooJ. Il contient plus de 25.000 entrées ; chaque entrée correspond à un emploi verbal associé à un ensemble de propriétés morphologiques (flexionnelles et dérivationnelles), syntaxiques (de structure et distributionnelles) et sémantiques (classe sémantique, synonymes).

En particulier, les constructions syntaxiques sont données systématiquement pour chaque emploi verbal. Les quatre grandes classes de constructions sont les classes A (constructions intransitives), N (constructions transitives indirectes), P (constructions pronominales) et T (constructions transitives directes). Ces constructions sont complétées par des informations distributionnelles sur le type des compléments (ex. Humain, non animé, etc.) et de prépositions utilisées (ex. *à*, *de*, etc.).

[1] Cf. (Silberztein 2003). NooJ est une plateforme de développement utilisée à la fois pour décrire les langues et pour construire des applications du TAL. NooJ est un logiciel gratuit et open source et est soutenu par l'initiative européenne Metashare et peut être téléchargé sur le site www.nooj4nlp.net.
[2] Cf. www.modyco.fr ; suivre la page « Ressources ». Le dictionnaire LVF est aussi accessible via le site WEB : http://rali.iro.umontreal.ca/rali/?q=fr/node/1237.
[3] Cf. (Dubois 1997).

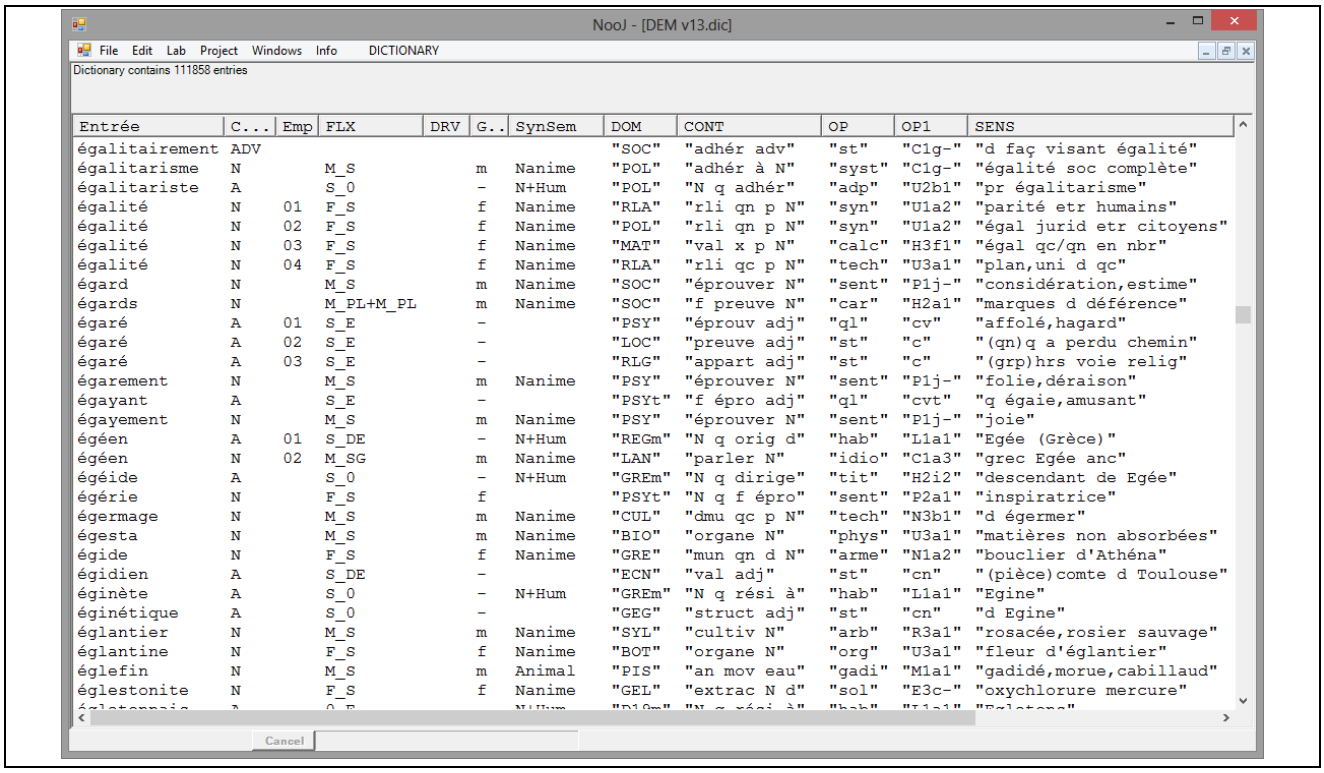
Par exemple, le code T1308 représente la structure syntaxique suivante :

Sujet humain (1), Verbe, Objet non animé (3), Complément instrumental (8)

(Silberztein 2010) décrivait l'implémentation du dictionnaire LVF ainsi que celle des grammaires génériques A, N, P et T dans la plateforme NooJ. Mais, faute d'information distributionnelle sur les noms, nous n'avions pas pu prendre en compte les informations distributionnelles caractérisant les actants de chaque emploi de LVF.

3 Le dictionnaire DEM

Le *Dictionnaire Electronique des Mots* (DEM) vient d'être publié par Jean Dubois et Françoise Dubois-Charlier[4]. Ce dictionnaire contient 145.135 entrées de toutes catégories, et se présente sous une forme similaire à celle du LVF.



Entrée	C...	Emp	FLX	DRV	G..	SynSem	DOM	CONT	OP	OP1	SENS
égalitairement	ADV						"SOC"	"adhér adv"	"st"	"C1g-"	"d faç visant égalité"
égalitarisme	N		M_S		m	Nanime	"POL"	"adhér à N"	"syst"	"C1g-"	"égalité soc complète"
égalitariste	A		S_0		-	N+Hum	"POL"	"N q adhér"	"adp"	"U2b1"	"pr égalitarisme"
égalité	N	01	F_S		f	Nanime	"RLA"	"rli qn p N"	"syn"	"U1a2"	"parité etr humains"
égalité	N	02	F_S		f	Nanime	"POL"	"rli qn p N"	"syn"	"U1a2"	"égal jurid etr citoyens"
égalité	N	03	F_S		f	Nanime	"MAT"	"val x p N"	"calc"	"H3f1"	"égal qc/qn en nbr"
égalité	N	04	F_S		f	Nanime	"RLA"	"rli qc p N"	"tech"	"U3a1"	"plan,uni d qc"
égard	N		M_S		m	Nanime	"SOC"	"éprouver N"	"sent"	"P1j-"	"considération,estime"
égards	N		M_PL+M_PL		m	Nanime	"SOC"	"f preuve N"	"car"	"H2a1"	"marques d déférence"
égaré	A	01	S_E		-		"PSY"	"éprouv adj"	"ql"	"cv"	"affolé,hagard"
égaré	A	02	S_E		-		"LOC"	"preuve adj"	"st"	"c"	"(qn)q a perdu chemin"
égaré	A	03	S_E		-		"RLG"	"appart adj"	"st"	"c"	"(grp)hrs voie relig"
égarement	N		M_S		m	Nanime	"PSY"	"éprouver N"	"sent"	"P1j-"	"folie,déraison"
égayant	A		S_E		-		"PSYt"	"f épro adj"	"ql"	"cvt"	"q égaie,amusant"
égayement	N		M_S		m	Nanime	"PSY"	"éprouver N"	"sent"	"P1j-"	"joie"
égéen	A	01	S_DE		-	N+Hum	"REGm"	"N q orig d"	"hab"	"L1a1"	"Egée (Grèce)"
égéen	N	02	M_SG		m	Nanime	"LAN"	"parler N"	"idio"	"C1a3"	"grec Egée anc"
égéide	A		S_0		-	N+Hum	"GREm"	"N q dirige"	"tit"	"H2i2"	"descendant de Egée"
égérie	N		F_S		f		"PSYt"	"N q f épro"	"sent"	"P2a1"	"inspiratrice"
égermage	N		M_S		m	Nanime	"CUL"	"dmu qc p N"	"tech"	"N3b1"	"d égermer"
égesta	N		M_S		m	Nanime	"BIO"	"organe N"	"phys"	"U3a1"	"matières non absorbées"
égide	N		F_S		f	Nanime	"GRE"	"mun qn d N"	"arme"	"N1a2"	"bouclier d'Athéna"
égidien	A		S_DE		-		"ECN"	"val adj"	"st"	"cn"	"(pièce)comte d Toulouse"
éginète	A		S_0		-	N+Hum	"GREm"	"N q résі à"	"hab"	"L1a1"	"Egine"
éginétique	A		S_0		-		"GEG"	"struct adj"	"st"	"cn"	"d Egine"
églantier	N		M_S		m	Nanime	"SYL"	"cultiv N"	"arb"	"R3a1"	"rosacée,rosier sauvage"
églantine	N		F_S		f	Nanime	"BOT"	"organe N"	"org"	"U3a1"	"fleur d'églantier"
églefin	N		M_S		m	Animal	"PIS"	"an mov eau"	"gadi"	"M1a1"	"gadidé,morue,cabillaud"
églestonite	N		F_S		f	Nanime	"GEL"	"extrac N d"	"sol"	"E3c-"	"oxychlorure mercure"

1. Le dictionnaire DEM

De ce dictionnaire, nous avons dans un premier temps exclus les locutions (mots composés et expressions figées), les mots grammaticaux ainsi que les verbes puisque ceux-ci sont déjà décrits dans le dictionnaire LVF. Le dictionnaire résultant contient donc 111.858 entrées lexicales. Nous avons donc entrepris de l'implémenter dans NooJ. Pour ce faire, nous avons dû associer à toutes les entrées concernées un modèle flexionnel. Nous avons pour cela utilisé les modèles flexionnels de (Trouilleux 2012)[5].

Le dictionnaire résultant, implémenté dans la plateforme NooJ, contient 82.192 noms[6], parmi lesquels figurent plus de 6.000 entrées lexicales qui ont les deux catégories Nom et Adjectif, par exemple *abolitionniste*. Le dictionnaire DEM,

[4] Cf. (Dubois 2010).
[5] Le dictionnaire DEM étant bien plus large que le dictionnaire DM, il a fallu décrire la flexion de plus de 50.000 nouvelles entrées ; merci à Denis Le Pesant pour son aide.
[6] Parmi les noms recensés dans le DEM, figurent un grand nombre d'entrées lexicales qui ont les deux catégories Nom et Adjectif, par exemple *abolitionniste*. Le dictionnaire DEM, contrairement à d'autres dictionnaires, ne dédouble donc pas les éléments du vocabulaire qui ont deux fonctions syntaxiques.

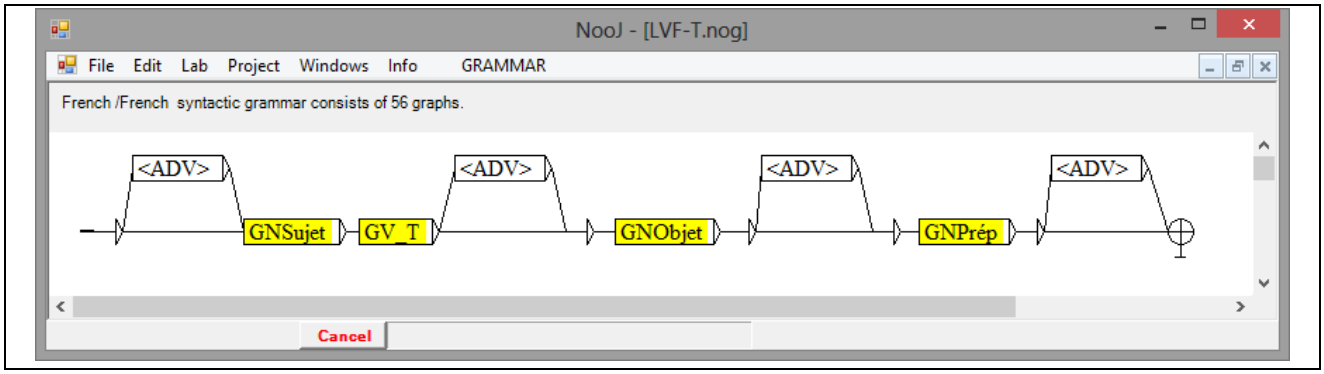
contrairement aux autres dictionnaires utilisés en TAL, ne dédouble pas les éléments du vocabulaire qui ont deux fonctions syntaxiques. Parmi les noms, plus de 15.000 ont été répertoriés comme humains.

4 Analyse syntaxique

Avec NooJ, on peut construire des grammaires syntaxiques pour reconnaître des phrases, puis les appliquer à des textes de taille importante. Les quatre types de phrases de base sont :

— A (constructions intransitives), ex. : *On arrive à Lyon*
— N (constructions transitives indirectes), ex. : *Les échecs alternent avec les succès*
— P (constructions pronominales), ex. : *On s'accommode de la situation*
— T (constructions transitives directes), ex. : *La chaleur accable les estivants*

Il suffit ensuite d'insérer chacun la grammaire de chaque schéma de phrase une grammaire des groupes nominaux telle que celle décrite par (Silberztein 2004)[7].



2. Grammaire T des phrases transitives

Si l'on dispose d'informations lexicales riches telles que celles du dictionnaire LVF, on peut les utiliser dans les grammaires afin d'éviter de nombreux résultats faux (« false positive »). Ainsi par exemple, les quatre types de constructions syntaxiques précédents sont systématiquement décrits dans le dictionnaire LVF : on peut donc associer chacune des quatre grammaires précédentes aux verbes qui acceptent les constructions correspondantes. Par exemple, utiliser des symboles NooJ comme <V+CONS="^T.*"> dans une grammaire permet de ne reconnaître que les verbes qui entrent dans les constructions de type T (transitives directes). On évite ainsi de reconnaître comme phrases transitives des phrases comme *Luc dort la nuit*. (Silberztein 2010) montrait comment construire des grammaires syntaxiques qui utilisent les données du dictionnaire LVF afin de départager les différents sens (ou emplois) des verbes.

Les codes de constructions associées aux entrées lexicales de LVF contiennent aussi des informations distributionnelles sur le sujet et les compléments de chaque emploi. Ainsi par exemple, pour le sujet du verbe :

1 : noms humains	2 : noms d'animaux	3 : noms de choses	4 : phrases	5 infinitives
6 : noms humains pluriel	7 : noms de choses pluriel	9 : noms concrets		

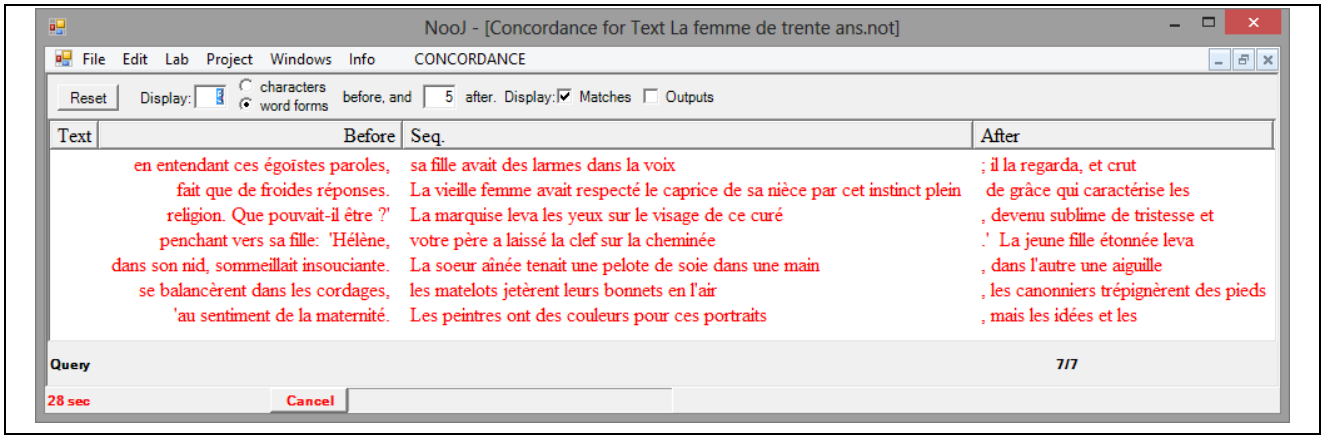
Grâce à l'intégration du DEM dans NooJ, on peut donc utiliser ces informations en les intégrant dans chacune des quatre grammaires syntaxiques A, N, P et T, simplement en associant les contraintes distributionnelles de LVF aux entrées lexicales du DEM. On peut donc construire des grammaires encore plus fines que celles décrites dans (Silberztein 2010), puisqu'on peut aussi vérifier que chaque emploi verbal a les « bons » types de sujet et de compléments.

[7] La grammaire des groupes nominaux est essentiellement l'implémentation de la grammaire des déterminants de (Gross 1977).

Ainsi par exemple, la construction transitive directe "T13..", sujet humain, complément d'objet direct de chose (que l'on trouve dans *Luc abaisse le rideau avec une manivelle*) est toujours traitée par la grammaire T ci-dessus, mais est maintenant associée aux deux contraintes distributionnelles sur les noms-têtes des groupes nominaux des grammaires GNSujet et GNObjet :

<N+Hum>/<VCONS="^T1">, <N+Nanime>/<VCONS="^T.3">

Le premier terme sélectionne un nom humain (<N+Hum>) si la construction associée au verbe (VCONS) a pour valeur une chaîne de caractères reconnue par l'expression rationnelle "^T1", qui signifie : le code de construction commence (^) par le caractère « T », suivi par le caractère « 1 ». Le second terme sélectionne un nom non-animé (<N+Nanime>) si la construction associée au verbe (VCONS) a pour valeur une chaîne de caractères reconnue par l'expression rationnelle "^T.3", i.e. le code de construction doit commencer par le caractère « T », peut être suivi par n'importe quel caractère (« . »), puis par le caractère « 3 ». En intégrant ces deux contraintes à la grammaire générique T et en appliquant celle-ci à des corpus de textes, on obtient des concordances comme la suivante :



3. Concordance sur la structure "^T13.*"

Appliquer à nos corpus de textes les grammaires prototypiques A, N, P et T en tenant compte des contraintes distributionnelles a permis d'améliorer considérablement la précision de la recherche, par rapport aux résultats décrits par (Silberztein 2010) : les erreurs ont toutes pour origine une confusion systématique entre compléments circonstanciels et compléments instrumentaux (le code 8 dans LVF). En revanche, les contraintes distributionnelles ont réduit le rappel, puisqu'il n'est plus possible de retrouver des constructions qui contiennent un pronom (par ex. *Il l'a abaissé avec cela*), et toutes les métonymies sont maintenant exclues (par ex. *La table a éclaté de rire* dans le sens de *Les personnes autour de la table ont éclaté de rire*). Mais résoudre les références et les métonymies ne fait pas partie du projet strictement linguistique : nous pensons donc, paradoxalement, qu'un logiciel de TAL qui bute sur ces problèmes constitute un progrès significatif pour la linguistique par rapport à un logiciel de TAL qui ne distinguerait pas de différence entre *Les étudiants ont éclaté de rire* et *La table a éclaté de rire*.

En conclusion, il est désormais possible d'extraire automatiquement d'un corpus les phrases qui contiennent un emploi (i.e. un sens) spécifique d'un verbe : aucun autre outil de linguistique de corpus ne permet de faire ce type d'opération ; il s'agit là aussi d'un saut qualitatif significatif pour la linguistique de corpus.

Références

DUBOIS JEAN, DUBOIS-CHARLIER FRANÇOISE, 1997. *Les Verbes français*. Paris : Larousse-Bordas.

DUBOIS JEAN, DUBOIS-CHARLIER FRANÇOISE, 2010. *Dictionnaire électronique des mots*.

GROSS MAURICE, 1977. *Grammaire transformationnelle du français, 2 : Syntaxe du nom*. Larousse : Paris.

SILBERZTEIN MAX, 2004. Une description formalisée des déterminants français. In *Hommage à la mémoire de Maurice Gross*. Linguisticae Investigationes, E. Laporte, C. Leclère, M. Piot, M. Silberztein Eds. pp. 589-600.

SILBERZTEIN MAX, 2005. NooJ Dictionaries. In *Proceedings of the 2nd Language and Technology Conference*. Poznan.

SILBERZTEIN MAX. 2010. La formalisation du dictionnaire LVF avec NooJ et ses applications pour l'analyse automatique de corpus. In *Théorie, empirie, exploitation : l'exemple des travaux de Jean Dubois sur les verbes français*. Langages n° 179-180, Danielle Leeman, Paul Sabatier Eds.