Morfetik : une ressource lexicale pour le TAL

Michel Mathieu-Colas

Morfetik : une ressource lexicale pour le TAL

Michel MATHIEU-COLAS

LDI (UMR 7187), CNRS et Université Paris 13

1 Introduction

Le traitement automatique des langues exige avant toute chose une reconnaissance précise des formes, ce qui présuppose un recensement lexical aussi rigoureux et complet que possible. Dans cette perspective, nous avons entrepris d'élaborer une ressource linguistique ayant les caractéristiques suivantes : large couverture, précision et fiabilité des informations, respect des normes, évolutivité.

Les données sont structurées sous forme de tables et servent de point de départ à un système de traitement (*Morfetik*) qui associe un moteur de flexion, un dictionnaire des formes fléchies, des interfaces de consultation et d'interrogation, ainsi qu'un ensemble d'outils permettant la maintenance et l'exploitation des ressources[1].

Nous présentons ici la composante linguistique du système pour le traitement des mots simples (un autre module consacré aux mots composés – ou plus largement à l'ensemble des unités polylexicales – est en voie d'achèvement).

2 Sources

Le recensement lexical a fait appel à de nombreuses sources lexicographiques. Pour ce qui est de la langue générale, les dictionnaires les plus courants ont été pris en compte (y compris des dictionnaires bilingues). En cas de désaccord (variantes graphiques), toutes les formes attestées ont été retenues.

Nous présentons ici les principales sources consultées :

- le *DELAS* (Dictionnaire électronique du LADL, cf. B. COURTOIS 1990) ;
- le *Petit* et le *Grand Robert* ;
- le *Petit Larousse illustré*, le *Lexis*, le *Grand Larousse encyclopédique* et le *Grand Dictionnaire encyclopédique Larousse* (GDEL) ;
- le *Trésor de la langue française* ;
- le *Harrap's* et le *Robert & Collins* ;
- des dictionnaires d'argot ;
- des tables de conjugaison (dont le *Bescherelle* et les *Verbes logiques* de A. DUGAS) ;
- *Le Bon Usage* de GREVISSE et des dictionnaires de « difficultés » pour le traitement des cas problématiques.

[1] Le développement et les applications informatiques sont assurés par plusieurs membres du LDI (notamment Pierre-André Buvet, Emmanuel Cartier, Ahmed Elchheb et Fabrice Issac), avec la collaboration de Yassine Madiouni.

S’agissant des termes spécialisés, l’exploration s’est étendue relativement loin. D’une part, des dictionnaires encyclopédiques ont été consultés : c’est ainsi qu’une partie non négligeable de la nomenclature du GDEL a été intégrée. D’autre part, certaines spécialités ont donné lieu à une recherche plus approfondie (par exemple la médecine et la minéralogie).

Au total, 106 884 mots simples ont d’ores et déjà été identifiés, ainsi répartis :

noms :	69950	pronoms :	68	prépositions :	58
adjectifs :	24405	verbes :	10232	conjonctions :	18
déterminants	59	adverbes :	1894	interjections :	200

Naturellement, l’inventaire n’est pas clos. Il devra se poursuivre par l’ajout de néologismes et l’intégration de nouvelles spécialités. En outre, la confrontation avec le vocabulaire du Web, rendue possible par les programmes liés à *Morfetik*, permettra de compléter les lacunes et d’enrichir la terminologie.

3 Description : tables de lemmes et tables de flexions

La structure des tables étant différente selon les catégories morphosyntaxiques, nous présentons cinq groupes distincts. Si, pour certains types de mots (par exemple les adverbes), un simple listage suffit, pour d’autres catégories – noms, adjectifs et verbes –, il convient d’élaborer deux grilles complémentaires : d'une part des tables de flexion permettant d’identifier et de coder tous les types flexionnels, d'autre part des tables attribuant à chaque lemme le code flexionnel correspondant. Ce sont ces tables qui seront ensuite utilisées par le moteur de flexion pour produire l'ensemble de toutes les formes fléchies.

3.1. *Les noms*

Le noyau de la table des lemmes se présente ainsi :

Lemme	Catgram	Flex
a	nm	00
aa	nm	00
aabam	nms	
aalénien	nms	
aarite	nf	01
abaca	nm	01
abacule	nm	01
abaisse	nf	01
abaissée	nf	01

Table 1 : extrait de la table des noms

Pour des raisons à la fois théoriques et pratiques, les noms n'ont pas été fléchis en genre[2] : il n'y a donc dans la nomenclature que des noms masculins et des noms féminins. Pour ce qui est du nombre, certains mots ont reçu une marque de *singulier obligatoire* (nms ou nfs) : le *bétail*, le *ça*, la *camarde*... Il arrive aussi que la forme au singulier coexiste avec une forme variable :

une *cloche*, des *cloches* / la *cloche* (=les clochards)
un *privé*, des *privés* (détective) / le *privé* (vs le *public*)
un *général*, des *généraux* / le *général* (vs le *particulier*)

Quant aux *pluriels obligatoires* (nmp ou nfp), ils correspondent à des situations diverses : singulier inexistant (*funérailles*), différence de genre (*orgues*, employé au féminin comme pluriel emphatique), problèmes d'homographie (*le frais* vs *les frais*). Quelques pluriels « sémantiques » ont été pris en compte : *assises* (vs *assise*), *ciseaux* (vs *ciseau*), *vacances* (vs *vacance*), etc. Mais la question reste ouverte quant au degré d'intégration de ces différences dans un dictionnaire morphologique.

En cas de variantes de casse, seule la forme avec minuscule figure en entrée (*angélus* plutôt que *Angélus*). Toutefois les deux graphies sont dégroupées quand elles correspondent à des différences d'emploi (*état/État, église/Église, basquaise/Basquaise*, etc.).

Deux champs supplémentaires permettent de prendre en charge certaines particularités graphiques. Les ligatures « œ » ou « æ » (par ex. *œuvre* ou *nævus*) font l'objet d'un champ spécifique. Il en va de même pour les graphies savantes (*ācrama, afición, chāfi`isme*) :

Lemme	**Catgram**	**Flex**	**GraphSav**	**Ligature**
acrama	nm	00	ācrama	
aficion	nfs		afición	
chafiisme	nm	01	chāfi`isme	
naevus	nm	00;1D		nævus
oeuvre	nf	01		œuvre

Table 1a : extrait de la table des noms

Pour les types de flexion, nous avons établi 59 codes, couvrant toutes les formes attestées, des plus fréquentes (00 : pluriel identique au singulier ; 01 : pluriel en –s ; etc.) aux plus rares. En particulier, 44 codes ont été consacrés aux emprunts, par exemple, pour les pluriels en –i :

[2] Nous avions dans un premier temps, à l'instar du *DELAS*, regroupé sous des entrées communes les noms à double genre (*un élève, une élève ; un instituteur, une institutrice*), en intégrant les marques de genre dans les codes de flexion. Ce choix impliquait une prise en compte des emplois, qu'il s'agisse de la distinction humain / non-humain (*laitier, ère*, n. [humain] ; *laitier*, nm [inanimé] ; *laitière*, nf [animal], cf. A. REY 1977 : 28) ou d'une polysémie interne aux humains (*maître, esse*, n. ; *maître*, nm ; *maîtresse*, nf). Mais ce traitement présente, dans le détail, de très nombreuses difficultés d'application : peut-on dire que *boulangère* soit simplement le féminin de *boulanger* (cf. GDEL : « Femme d'un boulanger, qui travaille à la boutique ») ? En outre, du point de vue théorique, le genre n'a pas le même statut pour les noms et les adjectifs (cf. I. MEL'ČUK 2000). Nous proposons, en conséquence, d'adopter au départ des entrées purement morphologiques (nm *vs* nf) et de traiter les relations de genre, compte tenu de leur complexité, dans un module spécifique.

Code	Rad	S	P	*exemples*
16	0	-	i	ricercar / ri
17	1	a	i	stotinka / i
18	1	e	i	ricercare / i
19	1	o	i	pizzicato / i
1A	1	u	i	leu / lei
1B	1	s	i	kouros / oi
1C	1	s	ï	kouros / oï
1D	2	us	i	nævus / i

Table 2 : extrait de la table de flexion des noms

Le champ « Rad » indique le nombre de caractères à enlever pour construire un radical artificiel utilisé par le fléchisseur pour générer les formes fléchies.

En cas de flexions multiples, tous les codes sont indiqués sans distinguer, du point de vue morphologique, les nuances polysémiques (des *ciels* / des *cieux*) et les simples variantes (des *barmans* ou des *barmen*) ; seuls les véritables homonymes donnent lieu à des lignes distinctes (des *sols* vs des *sol* [note de musique]).

3.2. *Les adjectifs*

Les mêmes principes, pour l'essentiel, s'appliquent à la table des adjectifs :

Lemme	Catgram	Flex
aalénien	adj	50
ababouiné	adj	42
abactériémique	adj	40
abactérien	adj	50
abaissable	adj	40
abaissant	adj	42
abaissé	adj	42

Table 3 : extrait de la table des adjectifs

Le codage des flexions est ici plus complexe puisqu'il faut compter avec la flexion en genre, soit, en général, une possibilité de quatre formes différentes (*aalénien, -ienne, -iens, -iennes*) :

Code	Rad	MS	MP	FS	FP	*exemples*
42	0	-	s	e	es	petit
43	0	-	s	ë	ës	aigu
44	1	û	us	ue	ues	dû
45	0	-	s	he	hes	franc
46	2	ec	ecs	èche	èches	sec
47	1	c	cs	que	ques	public
48	0	-	s	que	ques	grec

Table 4 : extrait de la table de flexion des adjectifs

Au total 52 codes ont été définis, auxquels s'ajoutent 7 codes conçus plus spécialement pour les adjectifs à genre fixe (par ex. *abaisseur*, adjm, ou *enceinte*, adjf). Les contraintes de nombre sont exceptionnelles (*opimes*, adjfp).

Si l'un des deux genres est particulièrement rare (au point de ne pas apparaître dans certains dictionnaires), nous mentionnons entre parenthèses la forme la plus fréquente : *gringalet* [adj(m)] s'emploie surtout au masculin, *accort* [adj(f)] est plus usuel au féminin.

3.3. *Déterminants et pronoms*

Nous avons introduit les sous-catégorisations suivantes :

D:Déf	Déterminant *défini*
D:Dém	Déterminant *démonstratif*
D:Ind	Déterminant *indéfini*
D:Int	Déterminant *interrogatif*
D:Excl	Déterminant *exclamatif*
D:Num	Déterminant *numérique*
D:Part	Déterminant *partitif*
D:Poss	Déterminant *possessif*
D:Rel	Déterminant *relatif*

P:Ddém	Pronom *démonstratif*
P:Ind	Pronom *indéfini*
P:Int	Pronom *interrogatif*
P:Pers	Pronom *personnel*
P:Poss	Pronom *possessif*
P:Rel	Pronom *relatif*

En raison de l'irrégularité des flexions, nous énumérons toutes les formes plutôt que d'introduire des codes spécifiques. Nous indiquons également, quand il y a lieu, les formes élidées et autres variantes :

Lemme	Forme	Catgram	Genre	Nombre	Personne
ce	c'	P:Dém			
ce	ç'	P:Dém			
ce	ce	P:Dém			
ce	ce	D:Dém	M	S	
ce	ces	D:Dém	M	P	
ce	ces	D:Dém	F	P	
ce	cet	D:Dém	M	S	
ce	cette	D:Dém	F	S	

Table 5 : extrait de la table des déterminants et pronoms

3.4. *Les verbes*

Le codage des verbes est évidemment beaucoup plus complexe. Chacun des codes associés aux lemmes (verbes à l'infinitif) correspond à un type de conjugaison spécifique qui doit être défini avec précision. La perspective d'un traitement automatique s'accommode mal des notes en bas de page et autres gloses marginales : cela signifie que toute variation dans le paradigme, la plus infime variante (*rassi* ou *rassis*), la moindre défectivité (*endormie* / **redormie* / ?*dormie*) détermine l'émergence d'un nouveau code.

Nous ne pouvons entrer ici dans le détail du codage. Indiquons seulement que, à côté des codes élémentaires et du code 000 pour les verbes sans flexion (infinitifs isolés : *quérir, férir, malfaire,* etc.), des sous-spécifications ont été prévues, principalement dans deux cas :

- Pour les verbes en E..ER (008) et en É..ER (012), l'adjonction de lettres au code reflète directement les consonnes intervocaliques (par ex. : 008m pour le type *semer*, 012s pour le type *léser*).
- Par ailleurs, une partie des formes peut être définie par simple « filtrage » d'un autre modèle, à l'aide de « points de défectivité ». Pour prendre l'exemple le plus simple, .z indique l'invariabilité du participe passé (*manger* est codé 001, mais *marcher* 001.z : **marchée*).

On aboutit ainsi à 222 codes, décrits dans une table comprenant les champs suivants :

CHAMPS	EXEMPLE
numéro de code	036
exemple-modèle	ACQUÉRIR
radical (nombre de caractères à soustraire de la forme canonique)	-4
radical-modèle	ACQU
désinences de l'infinitif	*érir*
désinences des formes conjuguées (45 champs)	*iers, iers, iert, érons...*
désinences des participes (5 champs)	*érant, is, ise, is, ises*

Table 6 : structure de la table de flexion des verbes

Les « radicaux » et « désinences » ainsi décrits ne correspondent pas nécessairement au découpage morphologique. Afin de faciliter le traitement automatique, le radical est défini comme le plus petit dénominateur commun : SER- pour SERVIR (pour construire *sers* et *sert*), V- pour VOULOIR (à cause de *veux* et *veut*), etc. A la limite, le radical peut être une forme vide (*être, avoir, aller*).

De manière complémentaire, les désinences s'ajoutent au radical pour construire les formes fléchies. Le signe = symbolise les désinences zéro (forme fléchie identique au radical, par ex. *vêt* ou *rend*). En cas d'inexistence d'une forme (défectivité), la désinence est remplacée par un tiret.

A partir de la table ainsi conçue, il est aisé de générer automatiquement des tableaux de conjugaison.

Variantes flexionnelles

Certains modèles intègrent des variantes (elles sont séparées par des points-virgules dans les tables de codes). Il peut s'agir d'une forme isolée :

ÉCLORE (100.1) : ind. prés. (3s) = *il éclôt* ou *il éclot*

d'un ensemble de formes :

> ASSEOIR (059) : *assois* ou *assieds*, *assoyais* ou *asseyais*, *assoirai* ou *assiérai*, etc. (avec mélanges possibles pour un même locuteur : *je m'assois, nous nous asseyons, ils s'assoient, assieds-toi*...)

ou d'un phénomène plus large englobant de nombreux verbes :

> futur et conditionnel des verbes en -É...ER (types 012<lettre>, 013 et 014 : CÉDER, CÉLÉBRER, RÉGNER, RÉVÉLER, RAPIÉCER, ASSIÉGER, etc.) ; si la tradition impose ici l'accent aigu (voir les guides de bon usage : *je céderai, il régnera*...), l'accent grave, plus conforme à la prononciation, tend à se généraliser (*je cèderai, il règnera*), au point d'être retenu comme seule graphie par la dernière édition du Dictionnaire de l'Académie.

Marquage des formes

Le signe « ° » indique les formes rares et/ou archaïques. Il permet de formaliser les indications éparses dans les dictionnaires et autres ouvrages de référence : *le plus souvent, en général, surtout, rarement,* etc. Du point de vue linguistique, la description y gagne en précision ; il existe des formes « semi-défectives », exceptionnelles mais attestées : comment décrire autrement les formes fléchies du participe de *dormir* (« las d'une nuit mal *dormie* », Boris Vian cité par GREVISSE dans *Le Bon Usage*) ? Sur le plan informatique, un tel système permet une double utilisation du lexique (standard / exhaustive), selon que les flexions marquées sont ou non prises en compte. Le choix n'est pas négligeable, compte tenu du nombre de formes en jeu.

Emplois

Il n'a pas été tenu compte, dans cette version, de la différenciation des emplois. Pour les rares verbes homonymes dont les flexions diffèrent, nous dégroupons les entrées, par ex. :

RALLER	« pousser un cri », en parlant du cerf (GR) : 001.z (*il ralle*)
	« aller de nouveau » (GDEL) : 015.1 (*il reva*...)
RESSORTIR	« sortir de nouveau » : 042 (*il ressort*)
	« être du ressort de » : 020.yz (*cette affaire ressortit à*...)

3.5. *Les mots « invariables »*

Les autres catégories – prépositions, adverbes, conjonctions, interjections – sont plus simples à décrire. Le champ « Forme » peut toutefois être mis à profit pour noter certaines variantes (notamment les élisions) :

Lemme	Forme	Catgram
jusque	jusqu’	Prép
jusque	jusque	Prép
jusque	jusques	Prép
que	qu’	C:Sub
que	que	C:Sub

ainsi que la variation exceptionnelle d'un adverbe comme *tout* (« elle était *toute* contente ») :

Lemme	Forme	CatGram
tout	tout	Adv
tout	toute	Adv
tout	toutes	Adv

4. Conclusion

Le système ainsi conçu permet de générer automatiquement l'ensemble des formes simples du français – environ 520 000 graphies correspondant à plus de 760 000 valeurs (compte tenu des homographies), en l'état actuel de la description. Ce travail est encore en cours puisque nous souhaitons y intégrer d'autres informations, qu'il s'agisse de la nomenclature (ajout de néologismes et de termes spécialisés) ou de la précision de la description :

- traitement approprié des noms à genre variable (voir *supra*, note 2) ;
- développement des marques relatives à la fréquence (possibilité de prise en compte d'indicateurs statistiques de présence en corpus) ;
- explicitation du lien entre variantes graphiques ;
- traitement plus fin des relations entre formes et emplois (voir l'exemple des pluriels « sémantiques »).

Nous préparons, dans le même esprit, un module consacré aux unités polylexicales (plus de 100 000 lemmes complexes).

Morfetik constitue ainsi un ensemble évolutif destiné à s'enrichir progressivement afin d'améliorer la chaîne de traitement des données textuelles.

BIBLIOGRAPHIE

BUVET Pierre-André, CARTIER Emmanuel, ISSAC Fabrice, MEJRI Salah (2007) : « Dictionnaires électroniques et étiquetage syntactico-sémantique », in HATHOUT Nabil, MULLER Philippe (eds), Actes des 14e Journées sur le Traitement Automatique des Langues Naturelles, pp. 239-248, Toulouse, IRIT Press.

COURTOIS Blandine (1990) : « Un système de dictionnaires électroniques pour les mots simples du français », *Langue française*, 87, Paris, Larousse, p. 11-22.

COURTOIS Blandine, SILBERZTEIN Max, eds (1990) : *Dictionnaires électroniques du français, Langue française*, 87, Paris, Larousse.

MATHIEU-COLAS Michel (1996-2006) : *Dictionnaire morphologique du français, I. Formes simples*, Rapport technique du LLI, Villetaneuse, Université Paris 13.

MEL'ČUK Igor (2000) : « Un FOU / une FOLLE : un lexème ou deux ? », in *Lexique, Syntaxe et Sémantique*, Mélanges offerts à Gaston Gross à l'occasion de son 60e anniversaire, *Bulag*, numéro hors série, Centre Lucien Tesnière, p. 95-106.

REY Alain (1977) : Images et modèles. Du dictionnaire à la lexicologie, Paris, A. Colin.