Dictionnaires morphologiques du français contemporain : présentation de Morfetik, éléments d'un modèle pour le TAL

Aude Grezka[1] Emmanuel Cartier[2] Michel Mathieu-Colas[1]
(1) LDI UMR 7187, Université Paris 13 Sorbonne Paris Cité
(2) LIPN-RCLN UMR 7030, Université Paris 13 Sorbonne Paris Cité
aude.grezka@ldi.univ-paris13.fr, emmanuel.cartier@lipn.univ-paris13.fr, michel.mathieu-colas@univ-paris13.fr

Résumé. Dans cet article, nous présentons une ressource linguistique, Morfetik, développée au LDI. Après avoir présenté le modèle sous-jacent et spécifié les modalités de sa construction, nous comparons cette ressource avec d'autres ressources du français : le GLAFF, le LEFF, Morphalou et Dicolecte. Nous étudions ensuite la couverture lexicale de ces dictionnaires sur trois corpus, le *Wikipedia* français, la version française de *Wacky* et les dix ans du *Monde*. Nous concluons par un programme de travail permettant de mettre à jour de façon continue la ressource lexicographique du point de vue des formes linguistiques, en connectant la ressource à un corpus continu.

Abstract.
French Contemporary Morphological Dictionaries : Morfetik Database, Elements of a Model for Computational Linguistics
In this article, we present a morphological linguistic resource for Contemporary French called Morfetik. We first detail its composition, features and coverage. We compare it to other available morphological dictionaries for French (GLAFF, LEFF, Morphalou and Dicolecte). We then study its coverage on big corpora (French *Wikipedia*, French version of *Wacky* and *Le Monde* 10 years). We conclude with a proposition for updating the dictionary by connecting the resource with a continuously live corpus.

Mots-clés: dictionnaire, morphologie, français, ressource linguistique, corpus
Keywords: dictionary, morphology, French language, linguistic resource, corpus

1 Morfetik, une ressource morphologique pour le TAL

La ressource lexicale Morfetik, développée au laboratoire LDI, est un dictionnaire morphologique des mots simples du français[1]. Nous présentons ici une mise à jour importante de la ressource présentée en 2009 (Buvet *et al*., 2009 ; Mathieu-Colas *et al*., 2009).

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 les dictionnaires bilingues : le *DELAS* (Dictionnaire électronique du LADL, *cf.* B. Courtois, 1990) ; le *Petit Robert* 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. Pour les termes spécialisés, l'exploration a été largement étendue. Des dictionnaires encyclopédiques ont été consultés : c'est ainsi qu'une partie non négligeable de la nomenclature du GDEL a été intégrée.

1.1 Mises à jour

Au total, 102 962 lemmes (noms, adjectifs, déterminants, pronoms, verbes, adverbes, prépositions, conjonctions, interjections) et 758 035 formes ont ainsi été identifiés. L'inventaire n'est pas clos puisque, actuellement, nous rentrons dans la ressource lexicale :

1/ L'ensemble des propositions des Rectifications orthographiques du français de 1990 (http://www.academie-francaise.fr/sites/academie-francaise.fr/files/rectifications_1990.pdf). Celles-ci ont pour objectif de rectifier l'orthographe de certains mots, sans pour autant constituer une réforme. Elles permettent notamment de lever

[1] Cette ressource est le résultat du travail d'une vingtaine d'années de collecte et de description, sous la direction de Michel Mathieu-Colas.

l'ambiguïté de l'orthographe de certains mots. Ces rectifications touchent entre 2 000 mots d'un dictionnaire d'usage courant qui en contient de 50 000 à 60 000 et plus de 5 000 mots si on prend en compte ceux qui sont rares et techniques. Nous ne prenons en compte pour le moment que les règles orthographiques relatives aux mots simples :

RÈGLES	**EXEMPLES**	
	ORTHOGRAPHE TRADITIONNELLE	**ORTHOGRAPHE RÉFORMÉE**
Un certain nombre de mots remplaceront le trait d'union par la soudure, notamment : - les mots composés de *contr(e)-* et *entr(e)-* - les mots composés de *extra-*, *infra-*, *intra-*, *ultra-* - les onomatopées - les mots d'origine étrangère - les mots composés avec des éléments « savants »	*contre-appel* *extra-terrestre* *tic-tac* *week-end* *agro-alimentaire*	*contrappel* *extraterrestre* *tictac* *weekend* *agroalimentaire*
Pour montrer la prononciation du *u*, le tréma est, dans les mots comportant : *-guë-* et *-guï-*, déplacé sur cette lettre *-geure-*, ainsi qu'avec le verbe *arguer*, rajouté à cette lettre	*aiguë, ambiguë* *ambiguïté* *gageure, arguer*	*aigüe, ambigüe* *ambigüité* *gageüre, argüer*
Au lieu de l'accent aigu, emploi de l'accent grave dans un certain nombre de mots et au futur et au conditionnel des verbes qui se conjuguent comme *céder*.	*événement* *je céderai*	*évènement* *je cèderai*
L'accent circonflexe disparaît sur *i* et *u*, mais est maintenu dans les terminaisons verbales du passé simple (1ere et 2e personnes du pluriel), l'imparfait et le plus-que-parfait du subjonctif (3e personne du singulier) et en cas d'homonymie.	*coût* *entraîner,* *nous entraînons* *paraître, il paraît*	*cout* *entrainer,* *nous entrainons* *paraitre, il parait*
Les verbes en *-eler* ou *-eter* se conjuguent comme *peler* ou *acheter*. Les dérivés en *-ment* suivent les verbes correspondants. Exceptions : *appeler*, *jeter* et leurs composés.	*j'amoncelle, amoncellement* *tu époussetteras*	*j'amoncèle, amoncèlement* *tu époussèteras*
Les mots en *-olle* et les verbes en *-otter* (et leurs dérivés) s'écrivent respectivement *-ole* et *-oter*. Exceptions : *colle*, *folle*, *molle* et les mots de la même famille qu'un nom en *-otte* (comme *botter*, de *botte*).	*corolle* *frisotter, frisottis*	*corole* *frisoter, frisotis*
Les mots empruntés forment leur pluriel comme les mots français et sont accentués conformément aux règles qui s'y appliquent. Exceptions : les mots ayant conservé une valeur de citation (comme *des mea culpa*).	*des länder* *des sandwiches* *revolver*	*des lands* *des sandwichs* *révolver*

TABLEAU 1 : SYNTHÈSE DE LA RÉFORME DE L'ORTHOGRAPHE DE 1990 (MOTS SIMPLES)

ll y a, en outre, plus d'une soixantaine de modifications orthographiques isolées. Ce sont des modifications sur des mots divers : par exemple *charriot* sur le modèle de *charrue*, *boursoufflement* (au lieu de *boursouflement*), *boursouffler* (au lieu de *boursoufler*), *boursoufflure* (au lieu de *boursouflure*), *cahutte* (au lieu de *cahute*), etc.

2/ Les mots de la base France Terme (http://www.culture.fr/franceterme). Cette base est consacrée aux termes recommandés au *Journal officiel de la République française*. Il regroupe un ensemble de termes de différents domaines scientifiques et techniques mais ne constitue en aucun cas un dictionnaire de langue générale : *édumétrie, psychométrie, innumérisme*, etc.

3/ Les correspondances masculin-féminin, notamment la féminisation des noms de métier (un ou une *pilote*, un *professeur*/une *professeure*).

4/ Les pluriels sémantiques (une *assise*, les *assises* ; un *ciseau*, des *ciseaux* ; un *échec*, les *échecs* ; un *papier*, les *papiers* ; la *vacance*, les *vacances*).

5/ Le vocabulaire spécialisé : médecine, minéralogie, etc. (*abstension*, *acanthite*…).

Par la suite, nous souhaitons également enrichir la base par les formes verbales composées (choix de l'auxiliaire, identification des verbes pronominaux) et les mots composés.

1.2 Structuration

La structure des tables étant différente selon les catégories morphosyntaxiques, nous avons mis en place cinq groupes distincts. Pour certains types de mots (comme les adverbes), un simple listage suffit. En revanche, pour d'autres catégories (noms, adjectifs et verbes), il convient d'élaborer deux tables complémentaires : (i) des tables de flexion pour identifier et coder tous les types flexionnels ; (ii) 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. Au total, 226 codes de flexion pour les verbes ont été définis, 59 pour les adjectifs et 63 pour les noms.

A titre d'exemple, nous présentons ici les encodages retenus pour les adjectifs. Dans ce cadre, la table des lemmes va comprendre, pour chaque lemme, un identifiant vers son code de flexion. Dans la table des flexions, on trouvera les différentes informations liées à chaque flexion, ainsi que la forme à ajouter. Les 59 codes à genre variable ont été définis, sur le modèle suivant :

Code	Rad	Masculin sing.	Masculin plur.	Féminin sing.	Féminin plur.	Exemples
30	0					albinos, ocre
31	0			e	es	gris
32	1	s	s	ce	ces	tiers
33	1	x	x	ce	ces	doux
34	1	x	x	se	ses	heureux
35	0			se	ses	gros
36	2	ès	ès	esse	esses	exprès
37	1	x	x	sse	sses	faux
38	1	s	s	te	tes	dissous
39	2	is	is	îche	îches	frais
3C	2	ux	ux	ille	illes	vieux
40	0		s		s	démocrate
42	0		S	e	es	petit

TABLEAU 2 : EXTRAIT DE LA TABLE DES FLEXIONS POUR LES ADJECTIFS

Ils sont précédés par quelques codes conçus plus spécialement pour les adjectifs à genre fixe :

Code	Rad	Masculin sing.	Masculin plur.	Féminin sing.	Féminin plur.	Exemples
00F	0	NULL	NULL			azygos
00M	0			NULL	NULL	preux
01F	0	NULL	NULL		s	enceinte
01M	0		s	NULL	NULL	extenseur
02M	0		x	NULL	NULL	bijumeau
03M	1	l	ux	NULL	NULL	multicanal
20M	2	an	en	NULL	NULL	gentleman

TABLEAU 3 : EXTRAIT DE LA TABLE DES FLEXIONS POUR LES ADJECTIFS

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

2 Couverture lexicale de la ressource

La ressource produite a été comparée avec les ressources lexicales analogues en français. La couverture lexicale a également été validée par comparaison avec trois corpus du français, les 10 ans du *Monde*, le *Wikipedia* français et la version française de *Wacky*.

2.1 Comparaison avec les ressources lexicales en français contemporain

Quatre autres ressources sont disponibles aujourd'hui[2] : le GLAFF, le Lefff, Morphalou et le Dicolecte. Nous donnons dans le tableau 4 les données principales pour ces différents dictionnaires[3].

	MORFETIK		GLAFF		MORPHALOU		LEFFF		DICOLECTE	
	lemmes	**formes**	**lemmes**	**formes**	**lemmes**	**formes**	**lemmes**	**formes**	**lemmes**	**formes**
ADJ	24 391	96 964	**42 204**	**125 409**	15 208	47 392	17 416	60 044	11 403	31 859
ADV	1 897	1 897	2 648	2 649	1 579	1 597	**3 119**	**3 143**	2 097	2 098
FCTW[4]	351	483	142	542	352	478	220	459	**3 727**	**3 783**
NC	66 393	138 963	**104 218**	**192 386**	41 000	80 261	40 109	84 276	44 139	98 532
PREP	57	60	50	56	(FCTW)		**128**	**159**	62	62
V	10 223	519 668	**21 402**	**1 085 422**	7 207	278 944	7 795	341 528	7 990	334 681
totaux	102 962	758 035	**170 664**	**1 406 464**	65 346	408 672	68 787	489 609	69 418	471 015

TABLEAU 4 : COMPOSITION DES DIFFÉRENTS DICTIONNAIRES MORPHOLOGIQUES

Le GLAFF (Hathout *et al*., 2014 ; Sajous *et al.,* 2013, 2014) est un dictionnaire extrait automatiquement à partir du Wiktionnaire français. Il comprend, pour chaque forme, les informations suivantes : la forme graphique, la description morphosyntaxique au format GRACE, le lemme, la ou les prononciation(s) en API et les prononciations équivalentes dans le format SAMPA. Il est à noter que le Wiktionnaire comprend un très grand nombre de gentilés et de lexies spécialisées, ce qui explique le très grand nombre de lemmes et d'entrées. Chaque entrée comprend également sa fréquence dans différents corpus (Wikipedia, LM10 et FrWac).

Le Lefff (Clément *et al*., 2004 ; Sagot, 2010) se place dans le modèle lexical Alexina, avec pour objectif d'être indépendant des langues spécifiques ainsi que des formalismes syntaxiques ; le format est compatible avec LMF (Francopoulo *et al.,* 2006) qui couvre les niveaux morphologique et syntaxique. Au niveau morphologique, chaque forme comprend son lemme, sa partie du discours et sa classe flexionnelle. Les classes flexionnelles sont définies dans le même esprit que celles de Morfetik. Les données elles-mêmes proviennent de plusieurs sources : récupération automatique (avec validation manuelle) à partir de techniques statistiques sur gros corpus, ainsi que récupération de données provenant d'autres ressources (essentiellement Multext, Veronis, 1998). Les noms propres, initialement intégrés à Lefff, ont ensuite été retirés. C'est la version révisée que nous prenons en compte ici.

Morphalou, développé par l'ATILF, est un lexique des formes fléchies du français construit à partir de la nomenclature du *Trésor de la Langue Française* (539 413 formes fléchies, pour 68 075 lemmes). Le dictionnaire résultant comprend un grand nombre de champs répondant à la norme LMF.

Dicolecte est un dictionnaire construit collaborativement pour les applications Open Office. Il comprend les informations suivantes : forme fléchie, lemme, étiquette grammaticale, métagraphe et métaphone, ainsi que des informations de fréquence dans trois corpus (Google 1-grams, Wikipedia, corpus de littérature issue du site gutenberg.org).

Le tableau 4 montre que, globalement, la couverture lexicale de Morfetik, du point de vue des lemmes comme des formes, est plus importante que celle des trois dictionnaires Morphalou, Lefff et Dicolecte, mais bien moindre que celle de GLAFF. Mais cette différence doit être affinée pour deux raisons principales : d'une part, le GLAFF comprend un grand nombre de formes dont la seule variation est la casse (première lettre en majuscule ou non, exemple : Aïd, aïd) ; d'autre part, un très grand nombre (1 071 327 lexies) ont une fréquence nulle dans les trois corpus que nous étudions, ce qui laisse 335 530 lexies « utiles » et 235 388 formes uniques. Les lexies à valeur nulle sont essentiellement des dérivés de noms propres (gentilés). Notons également que le GLAFF, pour les prépositions, adverbes et autres mots-outils, ne propose pas les listes les plus complètes.

Une autre différence entre les dictionnaires concerne le mode de description des variantes morphologiques : en effet, seuls Morfetik et le Lefff proposent des matrices morphologiques, les autres (Morphalou, DicoLecte, Glaff) se contentant de décrire les différentes formes liées à un lemme. Ces matrices sont particulièrement utiles car elles permettent d'étendre la couverture des dictionnaires de manière dynamique, notamment pour les parties du discours lexicales qui constituent des classes ouvertes. Nous allons voir dans la comparaison des dictionnaires sur corpus que ces matrices permettent une reconnaissance dynamique de formes inconnues, sans avoir à décrire les formes effectives.

Recouvrement lexicographique : les différents dictionnaires ont chacun des spécificités, et il convient à ce point d'étudier le recouvrement des dictionnaires, pour chacune des parties du discours, en partant du principe que les entrées des dictionnaires sont toutes valides. Le tableau 5 compare les trois dictionnaires les plus couvrants et explicites : les entrées communes (intersection), la combinaison des entrées (union), les entrées spécifiques à chaque dictionnaire, et les entrées présentes dans l'un des dictionnaires sauf Morfetik pour les verbes, noms et adjectifs. On constate que : 1/ l'intersection est faible (inférieure à 50% par rapport au dictionnaire le plus couvrant), et corrélativement l'union

[2] Le Delas fait aussi partie de cette liste, mais la comparaison a déjà été faite dans (Buvet *et al*., 2009). On consultera (Cougnon et Fairon, 2009) pour une mise à jour de cette ressource.

[3] En gras les volumétries les plus importantes.

[4] (FunctionWord) Correspond à : conjonction, pronom, déterminant, interjection.

améliore significativement la couverture ; 2/ les lexies spécifiques à chaque dictionnaire sont en nombre conséquent, notamment pour les noms et les adjectifs ; 3/ Morfetik : le nombre de lemmes manquants, présents dans au moins l'un des deux autres dictionnaires, est très important, mais il faut analyser ces « manques » ; en effet, parmi les 66 686 noms manquants, 64 153 sont des dérivés par affixation[5], ce qui laisse 2 545 lemmes manquants ; parmi les 11 877 lemmes verbaux, seul 2 – *voilà* - n'est pas un dérivé ; enfin, parmi les 35 491 lemmes adjectivaux, 6 178 sont des participes passés considérés comme adjectifs (GLAFF) et 29 027 sont des dérivés, ce qui laisse 286 lemmes manquants. Parmi les lemmes manquants, la totalité des lemmes proviennent du GLAFFF, dont une très grande majorité sont des emprunts récents, et, de fait, néologiques (exemples : *cokney, sabaoth, mamelouk, glamour...*), des termes techniques ou populaires (exemples : *sextil, tapuscrit, cornecul, pignouf, feuj, capout...*), ou encore comportent des erreurs typographiques (*succint, ...*). Somme toute, ces résultats nous semblent d'une part montrer l'intérêt de combiner les différents dictionnaires, et surtout de prévoir, en complément d'un dictionnaire des lexies usuelles, des matrices morphologiques permettant de reconnaître dans les textes des entrées liées à la productivité dérivationnelle. Cela apparaîtra encore plus clairement dans la confrontation des dictionnaires avec des corpus contemporains.

	NOMS	%	VERBES	%	ADJECTIFS	%
Entrées Morfetik	66 393		10 223		24 391	
Entrées Glaff	104 218		21 402		42 204	
Entrées Lefff	40 108		7 795		17 416	
Intersection entre les 3 diction.	31 473		6 856		8 614	
Union des entrées	133 079		22 100		59 882	
Lexies spécifiques Morfetik	21 610	32,55%	380	3,72%	9 106	37,33%
Lexies spécifiques Glaff	62 123	59,61%	10 966	51,24%	27 756	65,77%
Lexies spécifiques Lefff	3 179	7,93%	290	3,72%	7 505	43,09%
Autre dico sauf Morfetik	66 686		11 877		35 491	

TABLEAU 5 : COMPARAISON DES ENTRÉES (LEMMES) DES TROIS DICTIONNAIRES LES PLUS COUVRANTS

2.2 Couverture des dictionnaires sur corpus

Pour vérifier la couverture sur corpus, nous avons repris la méthodologie de Sajous (2014), en utilisant trois corpus suffisamment volumineux et représentatifs de la langue générale : version française de Wikipedia (août 2014, 100 millions de mots), version française Wacky[6] (1 milliard de mots), corpus des 10 ans du *Monde*, 1992-2002 (126 millions de mots). Nous avons étudié la couverture des trois dictionnaires les plus couvrants de la phase précédente (GLAFF, Morfetik, Lefff). Nous avons effectué un prétraitement des corpus afin d'éliminer l'effet « noms propres » (qui représentent près de 50% des lexies et constituent une classe ouverte non couverte par les dictionnaires morphologiques) en remplaçant dans les corpus toute entrée commençant par une majuscule, sauf premier mot de phrase n'ayant aucune autre occurrence commençant par majuscule, par la mention NP, sans en tenir compte dans les comptages. Nous avons également centré l'analyse sur les seules lexies simples.
Le tableau 6 présente les résultats, en effectuant la comparaison, d'une part, en ne considérant que les formes uniques (total formes uniques non reconnues), puis en considérant les occurrences et les fréquences, en faisant varier ce dernier paramètre : Fréquence >0 (toutes les occurrences du corpus), >1 (les formes ayant une fréquence supérieure à 1), etc. Pour chaque paramètre, nous notons le nombre d'occurrences non reconnues (exemple : FR Wikipedia – GLAFF – Fréquence > 100 : 2193 occurrences non reconnues) et le pourcentage par rapport à la totalité des occurrences du corpus. Les lignes COMBI correspondent à un dictionnaire construit par combinaison des entrées des trois dictionnaires.

Corpus	**Dictionnaire**	**Total formes uniques non reconnues**	**Total occurrences non reconnues (% des formes uniques)**					
			Fréquence >0	**Fréquence >1**	**Fréquence >4**	**Fréquence >9**	**Fréquence >100**	**Fréquence >1000**
FR Wikipedia (567 429 formes uniques, 99 731 049 occurrences)	**Morfetik**	419 000 (73,84%)	2 984 594 (2,99%)	166 765 (0,16%)	62 268	30 522	2 506	204
	GLAFF	415 637 (73,24%)	3 147 019 (3,15%)	164 492 (0,16%)	60 636	29 145	2 193	174
	LEFFF	470 496 (82,91%)	19 060 668 (19,11%)	206 639 (0,20%)	92 840	55 523	11 986	2 497
	COMBI	377 862 (66,59%)	1 910 637 (1,91%)	137 653 (0,13%)	45 803	20 596	1 134	88
FrWac (1 606 069 formes uniques, 1 031 810 340 occurrences)	**Morfetik**	1 378 805 (85,84%)	26 309 234 (2,55%)	657 668 (0,06%)	274 464	152 077	21 606	2 409
	GLAFF	1 367 236 (85,12%)	26 698 724 (2,58%)	651 172 (0,06%)	271 165	149 684	20 025	1 944
	LEFFF	1 474 666 (91,81%)	182 580 835 (17,69%)	739 433 (0,07%)	340 428	207 376	48 482	12 084

[5] Nous avons calculé le nombre des dérivés en utilisant des listes de préfixes et suffixes productifs en français, en considérant que si un lemme inconnu débutait, se terminait ou débutait et se terminait par l'un d'eux, il s'agissait d'un dérivé.

[6] http://wacky.sslmit.unibo.it/doku.php?id=corpora

	COMBI	1 320 418 (82,21%)	16 582 686 (1,60%)	614 572 (0,05%)	246 326	132 013	16 153	1 432
LM10 (227 262 formes uniques, 126 729 329 occurrences)	**Morfetik**	91 761 (40,37%)	721 952 (5,70%)	30 511 (0,02%)	10 393	4 924	533	63
	GLAFF	89 051 (39,18%)	6 205 818 (4,89%)	29 259 (0,02%)	9 815	4 557	496	85
	LEFFF	147 411 (64,86%)	26 220 353 (20,69%)	75 429 (0,05%)	44 928	32 917	12 118	2 918
	COMBI	78 761 (34,65%)	**352 674 (0,27%)**	23 301 (0,01%)	6 866	2 901	248	30

TABLEAU 6 : COUVERTURE DE GLAFF, MORFETIK, LEFF ET COMBINAISON DES TROIS SUR GROS CORPUS

Plusieurs enseignements en découlent :
1/ Le nombre de formes uniques inconnues se situe entre 82% (Lefff) et 73% (Glaff et Morfetik) du vocabulaire dans Wikipedia, entre 91% (Lefff) et 85% (Glaff et Morfetik) pour FrWac et entre 64% (Lefff) et 40% (Glaff et Morfetik) pour LM10 : la couverture du Lefff apparaît donc bien moindre que les deux autres, et Morfetik faisant jeu égal avec le Glaff malgré un lexique bien plus important pour Glaff. La disparité selon les corpus s'explique d'une part par les propriétés des deux premiers corpus, qui comprennent un très grand nombre de termes très spécialisés (Wikipedia essentiellement : *polyolefin, furocémide*, etc.), de noms propres sans majuscule initiale, d'erreurs typographiques (*techonopoles, accompagnéede, respectivemment*, etc.), de mots d'origine étrangère (*organizatsiya, roommate*,…). Les dix ans du *Monde* sont le corpus le plus « propre » de ce point de vue, mais révèlent également un nombre conséquent de néologismes, principalement par affixation (*supercentres, irremplaçabilité, autocommémore, juridictionnalisation*...), ainsi que des lexies composées dont les composants n'ont pas de valeur autonome (*statu quo, stricto sensu*...).
2/ Le nombre d'occurrences inconnues est dès le départ très faible (en dehors du Lefff), entre 3% (FrWikipedia, FrWac) et 5% (LM10), preuve d'une très bonne couverture lexicographique du Glaff et de Morfetik ; on notera que la moins bonne couverture concerne LM10, pourtant réputé comme corpus le plus proche d'un langage courant ; on notera également que si l'on ne considère que les formes ayant une fréquence supérieure à 1, les taux de couverture s'équilibrent (à environ 0,02%) pour tous les dictionnaires. Enfin, si l'on considère les formes inconnues ayant une fréquence supérieure à 10, il s'agit de lexies manquantes qui peuvent être utilement ajoutées aux dictionnaires ; ainsi, par exemple, Morfetik n'a pas pris en compte les abréviations des différentes unités de mesure (*km, cl*, …), les monnaies (*euro, yen*…).

3/ Pour chacun des corpus, la combinaison des lexiques conduit à une couverture plus grande, mais cet effet n'est plus visible dès que l'on considère les formes d'une fréquence supérieures à 1.

3 Conclusions et perspectives

Les dictionnaires morphologiques sont utiles pour l'analyse automatique des textes. Nous avons montré que Morfetik est la ressource la plus couvrante parmi les dictionnaires existants et qu'elle soutenait la comparaison avec le Glaff, la ressource collaborative, malgré une couverture certes moins grande, mais qui n'a qu'un effet limité si l'on considère l'exploitation de la ressource dans un système d'analyse des textes. Morfetik et ses mises à jour seront disponibles sous licence LGPL-LR à l'adresse suivante : http://extranet-ldi.univ-paris13.fr/Morfetik/

Enfin, pour être utilisé dans un système de TAL, un dictionnaire de formes n'est jamais suffisant, en raison de différents phénomènes discursifs et de la productivité continuelle des langues : un correcteur orthographique, un générateur de formes liées notamment à des matrices d'affixation permettant de rendre compte des dérivés, un traitement spécifique des noms propres et des termes sont ainsi indispensables. De ce point de vue, une étude complémentaire doit être menée afin d'exploiter les matrices morphologiques dont dispose Morfetik.

Les dictionnaires, comme toutes les ressources linguistiques, nécessitent également une confrontation continuelle avec des corpus. Du point de vue des formes linguistiques, cela revient à mettre en regard la ressource linguistique et un corpus continu, afin de suivre l'évolution fréquentielle des lexies, d'une part, de repérer les lexies qui sortent de l'usage (fréquence nulle sur une période), et celles qui semblent s'implanter (néologismes qui atteignent une fréquence suffisante, sur une période donnée). Un dictionnaire morphologique est donc l'un des composants d'un système plus large impliquant un corpus continu et un module néologismes, ainsi que différents outils pour suivre la fréquence d'usage des lexies du dictionnaire.

La combinaison des dictionnaires est également une piste intéressante pour améliorer la couverture sur corpus : nous avons montré l'intérêt d'une telle combinaison, chaque dictionnaire apportant des entrées spécifiques utiles à l'analyse automatique. La consolidation des ressources présentées ici sera prochainement proposée.

Enfin, un dictionnaire des unités linguistiques, pour être efficace en TAL, doit décrire un maximum d'unités polylexicales, même si cela nécessite des mécanismes de description incluant notamment les possibilités d'insertion entre composants, ainsi que des variations morphologiques des composants. Nous passons ainsi du dictionnaire au *construction*.

Références

BUVET P.-A., CARTIER E., ISSAC F., MATHIEU-COLAS M., MEJRI S., MADIOUNI Y. (2009). Morfetik, ressource lexicale pour le TAL, *TALN 2009*, Senlis, 24-26 juin 2009. <hal.archives-ouvertes.fr/halshs-00739036/>

CLÉMENT, L., LANG, B., SAGOT, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 1841–1844, Lisboa, Portugal.

COUGNON, L.-A., FAIRON, C. (2009). La mise à jour d'un dictionnaire électronique : Une expérience pédagogique liée à la mise à jour du Delaf, *Arena Romanistica*, 28th Conference on Lexis and Grammar, Bergen (29/09/2009-03/10/2009) - Vol. 1, no. 4, p. 58-71.

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

FRANCOPOULO G., MONTE G. (2006). *Lexical Markup Framework* (LMF aka ISO-24613), CD revision 9 : 15 mars 2006.

HATHOUT N., SAJOUS F., CALDERONE B. (2014). GLÀFF, a Large Versatile French Lexicon. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1007-1012, Reykjavik, Iceland.

MATHIEU-COLAS M. (2009). *Morfetik*, une ressource lexicale pour le TAL, *Cahiers de Lexicologie*, Paris, pp. 137-146.

SAGOT B. (2010). The *Lefff* , a freely available and large-coverage morphological and syntactic lexicon for French, In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 2744-2751, Istanbul, Turkey.

SAJOUS F., HATHOUT N., CALDERONE B. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2013)*, pp. 285-298, Les Sables d'Olonne, France.

SAJOUS F., HATHOUT N., CALDERONE B. (2014). Ne jetons pas le Wiktionnaire avec l'oripeau du Web ! Études et réalisations fondées sur le dictionnaire collaboratif. *Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014)*, pp. 663-680, Berlin, Allemagne.

VÉRONIS J. (1998). *Multext-Lexicons. A set of Electronic Lexicons for European Languages.* [CD-ROM]: Distributed by ELRA/ELDA.