Un dictionnaire et une grammaire de composés français

François Trouilleux
Clermont Université, Université Blaise-Pascal, EA 999, LRL
francois.trouilleux@univ-bpclermont.fr

Résumé. L'article présente deux ressources pour le TAL, distribuées sous licence GPL : un dictionnaire de mots composés français et une grammaire NooJ spécifiant un sous-ensemble des schémas de composés.

Abstract. The paper introduces two resources for NLP, available with a GPL license: a dictionary of French compound words and a NooJ grammar which specifies a subset of compound patterns.

Mots-clés : open source, ressources, dictionnaire, grammaire, mots composés
Keywords: open source, resources, dictionary, grammar, compound words

1 Introduction

Le logiciel NooJ (Silberztein, 2003) dispose d'un dictionnaire en source ouverte pour le français. Le présent article propose de le compléter par un dictionnaire de mots composés, qu'on appelle ici DM-C. Plusieurs dictionnaires électroniques utilisables pour le TAL sont accessibles gratuitement. Nous en avons observé cinq : le DELA (Silberztein, 1990), le Lefff (Sagot, 2010), Morphalou (Romary et al., 2004), Lexique (New, 2006) et Multext (Véronis, 1998). D'autres ressources électroniques pourraient être prises en compte, notamment le Wiktionnaire ou (Dubois et Dubois-Charlier, 2001), mais leur format (Wiki ou PDF) est moins directement exploitable [1]. M. Mathieu-Colas (1995a) a développé un dictionnaire formalisé contenant 13 000 formes à trait d'union, mais qui n'est, semble-t-il, pas diffusé.

Le tableau 1 ci-dessous permet d'introduire à la fois la problématique de ce travail et la solution que nous présentons ; il donne une comparaison quantitative des cinq dictionnaires considérés, pour les formes à trait d'union [2]. Les deux premières colonnes donnent le nombre de lemmes et de formes dans chaque dictionnaire. La colonne suivante (t. l.) donne le ratio nombre de lemmes / nombre de formes (« taux de lemmatisation »). Morphalou ne lemmatise pas du tout les composés ; le plus souvent, seule la forme du singulier ou l'infinitif figure dans le dictionnaire et quand plusieurs formes fléchies figurent dans le dictionnaire, chacune est son propre lemme. Le Lefff et Multext lemmatisent le mieux ; verbes mis à part, le taux de lemmatisation est de 1,76 pour ces deux dictionnaires. Le DELA et Lexique contiennent des verbes mais ne les fléchissent que très incomplètement.

Les six colonnes suivantes donnent pour chaque dictionnaire (par ligne) la part du total (en % du nombre de formes) qui est contenue dans chacun des autres (par colonne). Les chiffres de ces colonnes révèlent un taux de recouvrement assez faible entre les dictionnaires. Pour les compléter, notons que l'union des cinq dictionnaires compte 13 617 formes, dont seulement 4 344 (32 %) se retrouvent dans au moins deux dictionnaires et 569 (13 %) se retrouvent dans les cinq. Le DM-C intègre de 61 à 90 % des formes des cinq dictionnaires que nous avons observés. Pour des raisons qui apparaitront plus loin, le choix a été fait de ne pas viser un dictionnaire qui serait l'union des cinq dictionnaires de départ. Au DM-C s'ajoute une grammaire permettant de reconnaître différents types de composés. Pour chaque dictionnaire, la colonne G indique combien de mots de ce dictionnaire ne figurant pas dans le DM-C sont reconnus par la grammaire (en % du nombre de formes du dictionnaire). La colonne *total* fait la somme des deux colonnes

[1] Le Wiktionnaire a fait l'objet d'une conversion dans un format exploitable pour le TAL (GLAFF), mais les composés ne sont pas traités pour l'instant, cf. (Sajous *et al*., 2013).

[2] Par « forme », on entend le mot lui-même indépendamment de sa catégorie ou des traits qui lui sont associés. Les formes contenant un espace et un trait d'union (ex. *à rebrousse-poil*) ne sont pas comptées dans ce tableau (le DELA contient un nombre très important de composés avec espace, qui rendrait les croisements illisibles).

précédentes et indique donc le taux de couverture du couple DM-C + grammaire pour chaque dictionnaire. On obtient 99 % pour Morphalou car ce dictionnaire a servi de ressource centrale pour le DM-C. Le 1 % restant représente 46 formes volontairement exclues (essentiellement des erreurs).

	lemmes	formes	t. l.	dela	lefff	lex.	mph	mult.	dm-c	G	total
dela	3855	4997	1,3	100	29	33	29	29	61	24	85
lefff	1772	3307	1,87	44	100	49	39	58	89	6	95
lex.	3430	4243	1,24	38	39	100	35	42	68	25	93
mph	5608	5608	1	26	23	26	100	21	64	35	99
mult.	1658	4072	2,46	36	47	44	29	100	90	8	98

TABLE 1. Comparaison quantitative des dictionnaires pour les mots avec trait d'union

L'alinéa suivant présente des observations sur le contenu des cinq dictionnaires que nous avons retenus, qui justifient le développement de nos deux ressources. L'alinéa 3 présente notre grammaire NooJ pour un sous-ensemble de schémas de composition, et l'alinéa 4 notre dictionnaire et les solutions qu'il apporte aux problèmes observés.

2 Observation des dictionnaires

Le codage des composés en français fait l'objet de fluctuations bien connues, dont on retrouve la trace dans les dictionnaires électroniques. Le tableau 2 donne des exemples de formes différentes trouvées dans les cinq dictionnaires. Un premier axe de variation concerne l'usage du **trait d'union** *vs* **soudure** ou **espace**. N. Catach (1981) a montré que des dictionnaires imprimés variaient sensiblement sur ce point ; cette variation se retrouve dans les dictionnaires électroniques. Les exemples donnés dans les premières lignes du tableau 2 parlent d'eux-mêmes.

On sait aussi par ailleurs que les formes du **singulier** ou du **pluriel** de certains noms composés sont imprécises. Les rectifications de l'orthographe proposées en 1990 ont consacré cette variabilité en autorisant officiellement les variantes (cf. Catach, 1991). Le traitement des mots *accroche-cœur* et *monte-plats* illustre la variabilité à l'œuvre dans les dictionnaires (*s* et *p* indiquent le nombre associé, – indique l'absence de trait de nombre). Les noms *année-lumière* et *mot-valise* illustrent en outre des problèmes de couverture et de précision. Il est à noter que l'absence de marque de genre dans Morphalou ne signifie pas une invariabilité, mais plutôt une absence d'information sur les formes fléchies.

Un troisième axe de variation, moins souvent évoqué car plus spécifique aux dictionnaires pour le TAL, concerne la **segmentation** des unités lexicales. Si la lecture des entrées *plain-pied* et *stand-by* dans le TLF suggère que ces mots soient considérés comme des noms, l'expression *bras-le-corps* semble bien impossible sans la préposition *à*, ce qui doit conduire à ne retenir qu'une entrée adverbiale *à bras-le-corps*.

Enfin, on rencontre aussi des **variantes orthographiques**, telles que *laisser-passer* ou *medicine-ball*. En règle générale, ces variantes ne sont pas signalées par l'attribution d'un lemme commun aux deux formes, si bien que le lemme *laisser-passer* est compté comme figurant dans deux dictionnaires, alors que le concept, si l'on peut dire, est connu des cinq.

Les disparités relevées en terme de recouvrement des dictionnaires peuvent s'expliquer aussi par le fait que la composition est un procédé **productif**. Par exemple, Lexique contient *en propre* 88 lemmes associés à des formes avec un préfixe *re-* (p.ex. *re-contacter*, *re-belote*, mais aussi *re-café*, *re-drapeau*, *re-main*...), 163 avec *ex-*, 22 avec *co-*. Morphalou contient 123 formes commençant par *anti-* dont seulement 11 se retrouvent dans un autre dictionnaire. Le TLF, dont est dérivé Morphalou, donne des composés parfois comme mot vedette (p.ex. *presse-citron*), parfois à l'intérieur des articles (p.ex. *pubo-fémoral* figure dans l'article *fémoral*). En particulier, le TLF a dans sa nomenclature des préfixes, suffixes et autres « éléments formants », pour lesquels les articles donnent des listes de composés. Seulement 39,7 % des mots à trait d'union de Morphalou figurent comme mots vedettes dans le TLF.

L'objectif d'un dictionnaire classique comme le TLF est multiple : fournir un relevé des mots attestés, en donner la définition, en expliquer les emplois. S'agissant des composés, le premier de ces objectifs n'est pas central pour un dictionnaire destiné au TAL. Il est au moins aussi important d'avoir un système rendant compte des mécanismes de composition, qui permettra de détecter les nouvelles productions. On est tiraillé dans deux directions opposées : la lexicalisation d'un maximum d'unités et la reconnaissance des unités composées par des règles. C'est la raison pour laquelle le traitement que nous proposons se fait en partie par une grammaire, en partie par le dictionnaire.

DELA	***Lefff***	***Lexique***	***Morphalou***	***Multext***
contreplongée	*contre-plongée*	*contre-plongée*	*contreplongée*	
sociolinguistique	*sociolinguistique*		*sociolinguistique* *socio-linguistique*	*sociolinguistique*
cul de basse fosse		*cul-de-basse-fosse*	*cul de basse-fosse*	*cul-de-basse-fosse*
		pan bagnat		*pan-bagnat*
accroche-cœurs s *accroche-cœurs* p	*accroche-cœur* –	*accroche-cœur* s *accroche-cœurs* p	*accroche-cœur* –	*accroche-cœur* s *accroche-cœur* p *accroche-cœurs* p
monte-plats s *monte-plats* p	*monte-plat* p *monte-plats* –	*monte-plat* s *monte-plats* –	*monte-plats* –	*monte-plat* s *monte-plats* p
années-lumière p	*année-lumière* s *années-lumières* p	*année-lumière* s *années-lumière* p		
mot valise s	*mot-valise* s *mot-valises* p	*mot-valise* s	*mot-valise* –	
stand-by,N	*en stand-by*,ADV	*stand-by*,N	*stand-by*,N	*stand-by*,N
de plain pied,ADV *de plain-pied*,A		*plain-pied*,N	*plain-pied*,ADV *plain-pied*,N	*de plain-pied*,ADV
à bras-le-corps,ADV	*bras-le-corps,N*		*à bras-le-corps,ADV* *bras-le-corps,ADV*	
laissez-passer *laisser-passer*	*laissez-passer*	*laissez-passer*	*laissez-passer* *laisser-passer*	*laissez-passer*
		médecine-ball *medicine-ball*	*médecine-ball*	*médecine-ball* *medicine-ball*

TABLE 2. Exemples de disparités entre dictionnaires.

3 Une grammaire pour un sous-ensemble des composés

Définir un jeu de règles décrivant des schémas productifs de composition peut sembler a priori sans difficulté. M. Mathieu-Colas (1996), par exemple, donne une liste très détaillée de tels schémas. La difficulté apparaît cependant quand il s'agit de détecter les composés dans les textes *sans générer d'analyses incorrectes*. Nous présentons ici une grammaire qui porte sur un sous-ensemble de composés qui peuvent être identifiés syntaxiquement, hors contexte, avec une très haute précision, ce qui rend leur inclusion dans un dictionnaire dispensable.

Si on tente d'identifier les composés en présence d'un trait d'union, qu'on peut voir comme leur signature, on s'expose à deux problèmes : celui des juxtapositions (ex. *l'opposition consonnes-voyelles*) et surcompositions (ex. *un porte-filtre à café*), cf. (Mathieu-Colas, 1995b) et celui des ambiguïtés lexicales (p.ex. *ferme-auberge* ne désigne ni une auberge qui aurait une certaine fermeté, ni un instrument pour fermer les auberges). Etant donné ces difficultés, la lexicalisation des composés reste la meilleure solution pour les schémas N-N, N-A, A-A et V-N. En revanche, on peut identifier un sous-ensemble de schémas qui peuvent faire l'objet d'une analyse quasi déterministe. C'est ce que nous avons fait dans une grammaire NooJ, dont le tableau 3 donne les caractéristiques.

Quelques préfixes et quelques-uns des 53 verbes listés comme productifs pour le schéma V-N sont aussi des noms et attestés en position 1 d'un composé N-N ou N-A, comme dans p.ex. *auto-école, micro-cravate, photo-finish, serre-tunnel*... On a pris soin d'inclure ces suites N-N ou N-A dans le DM-C et on a inclus dans la grammaire, outre les schémas du tableau 3, deux règles permettant de construire ce type de suites avec ces formes ambigües. La grammaire reconnait donc par exemple *télé-surveillance* par deux schémas : PFX-N ou N-N.

Le schéma le plus productif dans cet ensemble, relativement aux dictionnaires, est de loin la composition avec préfixe (cf. tableau 4). D'autres schémas s'y apparentent : mot fonctionnel introducteur de syntagme (préposition ou numéral) ou modifieur antéposé (adjectif, adverbe ou point cardinal). La lexicalisation de 53 formes verbales permet en quelque sorte de transformer ces verbes en « préfixes », à la manière de la catégorie « élément de composition » du TLF. On rejoint donc dans cette grammaire la stratégie du *chunking*, qui cible les séquences précédant les têtes de syntagmes. Les schémas V-GP et N-GP relève quant à eux d'une stratégie de reconnaissance de la plus longue chaine : en présence d'un composé comme *Vaires-sur-Marne*, en corpus, il faut éviter de reconnaitre *sur-Marne* comme un composé PFX-N.

Schémas	***Exemples***
Composés sur préposition. Avec les prépositions *à*, *après*, *avant*, *en*, *hors*, *outre*, *pour* ou *sans*.	*à-côté, avant-programme, en-but, sans-opinon*
Composés avec préfixes. Les préfixes se combinent avec des mots de catégorie ADV, A, N ou V pour donner un mot de même catégorie. Le DM a été développé pour ce projet et contient 1371 préfixes.	*quasi-contractuellement, turbo-train, extra-plat, artério-veineux, sous-traiter, pré-enregistrer*
Composés sur verbe. Deux infinitifs, ou un verbe avec un adverbe antéposé, ou encore le schéma très productif V-N mais limité à un ensemble de 53 verbes identifiés comme les plus productifs [3].	*savoir-faire, bien-être, bien-aimé, attrape-couillon, cache-col, lave-linge, porte-bébé*
Composés numéral – nom pluriel.	*quatre-chevaux, trois-pièces*
Composés adjectif – nom. Schéma limité à 19 adjectifs attestés avec au moins trois occurrences dans un composé A-N du DM-C ou 6 autres adjectifs attestés et apparentés par le sens à l'un des 19 [4].	*blanc-seing, court-bouillon, double-toit, morte-saison, rouge-gorge, vif-argent*
Composés *sud*, *nord*, *est* ou *ouest* – adjectif.	*ouest-allemand, sud-vietnamien*
Composés verbe ou nom – groupe prépositionnel (V-GP et N-GP).	*tape-à-l'œil, abri-sous-roche*

TABLE 3. Types de composés identifiés par la grammaire.

4 Dictionnaire

Le DM-C a été construit en prenant Morphalou comme noyau central. Il contient en premier lieu les formes à trait d'union appartenant à au moins deux des cinq dictionnaires que nous avons examinés ou qui appartiennent à Morphalou seulement mais ne sont pas reconnues par notre grammaire. Le premier de ces deux critères permet de filtrer des erreurs. Le second critère permet de s'appuyer sur une ressource dont la licence permet clairement le réemploi et, pour les vérifications, sur les définitions du *Trésor de la langue française*, dont est extrait Morphalou. Le DM-C intègre également les lemmes composés avec espace de Morphalou (un peu moins de 300 lemmes), et les mots fonctionnels et quelques adverbes figurant déjà dans le DM des mots simples. On obtient un dictionnaire de 4910 lemmes.

4.1 Flexion des composés

On a vu dans l'introduction (cf. tableau TABLE 1) que Morphalou ne lemmatise pas les mots à trait d'union et contient peu de formes fléchies. C'est un des apports du DM-C que de fournir une flexion systématique des composés qu'il contient. Le formalisme de NooJ permet de définir des modèles de flexion auxquels on associe ensuite les entrées du dictionnaire. On a défini pour le DM-C 74 nouveaux modèles de flexion, qui s'ajoutent à l'ensemble déjà défini pour le DM.

Le pluriel des composés de type V-N et PREP-N a fait l'objet d'une réforme en 1990. Le DM-C intègre les recommandations de cette réforme, qui, rappelons-le, préconise des formes qui *s'ajoutent* à la graphie traditionnelle. Ainsi, les mots *accroche-cœur*, *monte-plat* et *taille-crayon* sont associés respectivement aux trois modèles de flexion qui suivent :

```
M_S_0     = <E>/m+s      |  s/m+p+Rec  |  <E>/m+p+Opt ;
M_0_S     = <E>/m+s+Rec  |  s/m+s+Opt  |    s/m+p     ;
M_0_S_0_S = <E>/m+s+Rec  |  s/m+s+Opt  |  <E>/m+p+Opt  |  s/m+p+Rec ;
```

Ces modèles génèrent deux pluriels pour *accroche-cœur*, deux singuliers pour *monte-plat* et deux singuliers, deux pluriels pour *taille-crayon*. 296 + 86 + 43 (78 %) des 544 composés de type V-N sont associés à ces trois modèles de flexion. Suivant une notation adoptée depuis la version 1.3 du DM, le trait `Rec` marque la graphie recommandée par la réforme et le trait `Opt` (« optionnel ») la graphie archaïsante. Le DM-C est à notre connaissance le premier dictionnaire pour le TAL à intégrer de façon systématique la flexion nouvelle des composés.

[3] Verbes bases d'au moins trois composés de type V-N, V-PRO (*brûle-tout*) ou V-GN (*trompe-la-mort*) dans Morphalou.

[4] Au moins 3 occurrences: *bas, beau, blanc, bon, court, double, faux, franc, grand, gros, haut, libre, mort, nu, petit, plat, plein, saint, tiers* ; apparentés : *long, rouge, vif, prime, quart, vert.*

4.2 Origine des composés et schémas de composition

Outre une flexion systématique, le DM-C donne deux informations supplémentaires : l'origine du mot ou son schéma de composition. Le DM-C contient 482 composés d'origine étrangère. Ces mots ont un attribut `Lang` dont la valeur est le code ISO 639 de la langue d'origine. La langue la plus fréquente est l'anglais : 58 % des mots d'origine étrangère, suivie du latin (30 %), de l'italien (5 %), puis de 19 langues différentes, chacune avec moins de 5 mots.

Les noms, verbes et adjectifs proprement français ont un attribut `Cmp` dont la valeur indique le schéma de composition. Le tableau 4 donne les fréquences de 43 schémas dont le code est obtenu par croisement des en-têtes de ligne et de colonnes. Ainsi, par exemple, 800 lemmes sont construits sur un schéma préfixe-nom (code PFX_N). A ces 43 codes s'ajoutent quelques autres codes, dont dCMP (dérivé de composé, ex. *court-circuiter*, 128 occ.), REP et QREP (répétition ou quasi-répétition, ex. *fric-frac*, 57 occ.), PH (composés par phrase, ex. *m'as-tu-vu*, 23 occ.).

	N	A	GP	PP	G	V	PRO	ADV	GN	SFX	NUM	P	total
PFX	800	513				35	1			10	1		1360
N	648	88	176	5	2			1	2				922
A	256	25	11	8	1	1							302
ADV	1	5	1	15	3	10						1	36
V	545	15	19			14	11	11	13		1		629
P	88	3				2	11		7		2		113
NUM	25		1								3		29
total	2363	649	208	28	6	62	23	12	22	10	7	1	3391

Table 4. Fréquence des types de composés.

4.3 Gestion des variantes

L'usage du trait d'union, comme on l'a rappelé au §2, est assez flottant. Le TLF contient des entrées dont la forme graphique inclut un trait d'union facultatif, telles que *saute(-)en(-)barque* ou *auto(-)féconder*. L'ambiguïté d'expansion des parenthèses – orthographe sans ou avec espace possible – a conduit les concepteurs de Morphalou à retenir seulement la version avec trait d'union. Morphalou a donc un biais en faveur des graphies avec trait d'union. On adopte pour le DM-C une approche qui tolère assez largement les alternances soudure/trait d'union et trait d'union/espace, grâce à deux caractères spéciaux fournis par NooJ : _ et =, qui sont interprétés respectivement comme « la chaine vide, un trait d'union ou un espace » et « un trait d'union ou un espace ». Ainsi, par exemple, une entrée `week_end,N` permet de reconnaitre à la fois *weekend*, *week-end* et *week end*.

Sont déclarés dans le DM-C avec le caractère _ les mots dont la soudure est recommandée par la réforme de 1990 (ex. *hotdog*) et les composés sur préfixe, à l'exception des préfixes *sous*, *demi*, *mi*, *semi*, *self*, *vice*, *ex* (signifiant « antérieurement »), des préfixes référant à des peuples (*franco*, *judéo*...) et des formes *non*, *quasi*, *arrière* et *social* que le DM catégorise aussi comme « préfixes ». On tient aussi compte d'exceptions telles que *extra-utérin* ou *contrexpertise*.

Sont déclarés avec le caractère =, de façon systématique, les noms composés de type N-A (*garde=champêtre*, *amour=propre*), N-N (*allocation=chômage*), N-GP (*bec=de=lièvre*) et V-N (*vide=grenier*), les cardinaux et ordinaux avec *et* (*vingt=et=unième*) et les mots étrangers (quand ils n'admettent pas aussi la soudure ; *pater=familias*, *pan=bagnat*). S'y ajoutent certains noms A-N, et certains adverbes ou adjectifs (ex. *bien=pensant*, *dos=à=dos*).

Les composés sur préposition, numéral, point cardinal ou par phrase sont déclarés avec trait d'union et sans alternative, à quelques exceptions près (p.ex. on admet *je ne sais quoi* et *qu'en dira-t-on*).

On sait qu'il y a des tendances fortes dans l'usage du trait d'union ; ainsi par exemple, il est la norme pour les composés V-N. Cette norme est cependant loin d'être appliquée systématiquement (cf. p.ex. vide-greniers.org), et elle est bien souvent inutile. Qu'on écrive *garde-champêtre* ou *garde champêtre*, ces deux séquences peuvent en confiance être analysées comme des réalisations d'un même composé. Il y a des cas où, certes, la présence du trait d'union est censée signer l'emploi d'un composé par opposition à un sens littéral. C'est le cas par exemple de l'adverbe *sur-le-champ* ou de noms comme *queue-de-pie*. Pour ces cas, le DM-C contient deux entrées : la version avec trait d'union qui sera analysée de façon déterministe comme un composé (via un trait spécial de NooJ : `+UNAMB`) et une version avec espaces qui sera analysée de façon non déterministe, à la fois comme une occurrence possible du composé ou comme une séquence de chacun des mots composant la suite. Nous faisons l'hypothèse (dont la validation demanderait un travail spécifique) que

l'emploi « fautif », avec espaces, des expressions à traits d'union est en règle générale plus fréquent que l'emploi littéral des expressions correspondantes. Signalons cependant que si, pour une meilleure analyse des textes réels, le DM-C tolère la variation au niveau des formes, il propose une graphie précise au niveau des lemmes.

5 Conclusion

Le DM-C et la grammaire NooJ présentés ici sont téléchargeables sur le site du LRL et utilisables dans les termes de la licence GPL : http://lrl.univ-bpclermont.fr/spip.php?rubrique48. Avec ces deux ressources, on pense améliorer la couverture lexicale des composés, et apporter des éléments de solution au problème de la variabilité des usages de flexion et du trait d'union, ainsi qu'à celui de la productivité des schémas de composition. Il s'agit d'*éléments* de solution, mais circonscrire un espace où une analyse fiable peut être menée est en soi un résultat. 4035 formes des dictionnaires DELA, Lefff, Lexique et Multext réunis sont inconnues du DM-C. La grammaire en analyse 2682 (66 %). L'examen de ces analyses ne révèle que 7 erreurs : l'abréviation *c-à-d* comme N-GP, *avant-gauche* comme PREP-N et cinq composés sur des adjectifs de couleurs vus comme des composés A-N (ex. *blanc-bleu*). L'évaluation sur corpus des deux ressources serait un travail de recherche en soi. Elle nécessiterait de prendre en compte des variations selon les types de textes et les époques, susceptibles de mettre en cause la présence de certains composés dans le dictionnaire. Nous espérons que nos deux ressources pourront servir à mieux cerner ce qui, dans la composition, relève de la grammaire ou de la lexicalisation.

Références

CATACH N. (1981). *Orthographe et lexicographie. Les mots composés.* Paris : Nathan.

CATACH N. (1991). *L'orthographe en débat.* Paris : Nathan.

COURTOIS B., SILBERZTEIN M. (1990). *Dictionnaires électronique du français. Langue française* 87. Paris : Larousse.

DUBOIS J., DUBOIS-CHARLIER F. (2001). Composition et préfixation en français. Aix-en-Provence : chez les auteurs.

MATHIEU-COLAS M. (1995a). Un dictionnaire électronique des mots à trait d'union. *Langue française* 108, 76-85. Paris : Larousse.

MATHIEU-COLAS M. (1995b). « Syntaxe du trait d'union : Structures complexes ». *Linguisticæ Investigationes* XIX:1, 153-171. Amsterdam : John Benjamins B.V.

MATHIEU-COLAS M. (1996). « Essai de typologie des noms composés français ». *Cahiers de lexicologie* 69, 71-125.

NEW B. (2006). « Lexique 3 : Une nouvelle base de données lexicales. » *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006)*, Louvain, Belgique. http://www.lexique.org/ (version 3.80)

ROMARY L., SALMON-ALT S., FRANCOPOULO G. (2004). "Standards going concrete: from LMF to Morphalou". *Workshop on Electronic Dictionaries*, Coling 2004, Geneva.
www.cnrtl.fr/lexiques/morphalou/ (ATILF/Nancy Université - CNRS)

SAGOT B. (2010). "The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French". *7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malte.
http://alpage.inria.fr/~sagot/lefff.html (version extensionnelle 3.0)

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).*

SILBERZTEIN M. (1990). « Le dictionnaire électronique des mots composés. » *Langue française* 87, 71-83. Paris : Larousse. http://infolingu.univ-mlv.fr/ > *Données linguistiques* > *Dictionnaire* > *Téléchargement*

SILBERZTEIN M. (2003). NooJ Manual. http://www.nooj4nlp.net.

VÉRONIS J. (1998). *Multext-Lexicons. A set of Electronic Lexicons for European Languages*. ELRA Catalogue (http://catalog.elra.info), MULTEXT Lexicons, ref.: ELRA-L0010.