



Un dictionnaire et une grammaire de composés français

François Trouilleux

► **To cite this version:**

François Trouilleux. Un dictionnaire et une grammaire de composés français. 21e Traitement automatique des langues naturelles - TALN 2014, Jul 2014, Marseille, France. 2014. <hal-01005524>

HAL Id: hal-01005524

<https://hal.archives-ouvertes.fr/hal-01005524>

Submitted on 12 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 présent article s'inscrit dans un programme global de développement de ressources pour l'analyse du français avec le logiciel NooJ (Silberztein, 2003). On dispose depuis (Trouilleux, 2012) d'un dictionnaire de mots simples, le DM, qu'on complète ici 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, les dictionnaires de l'Académie française mis en ligne par le CNTRL ou (Dubois et Dubois-Charlier, 2001), mais leur format (Wiki, HTML ou PDF) est moins directement exploitable¹. M. Mathieu-Colas (1995a) a développé un dictionnaire formalisé contenant 13 000 formes à trait d'union, mais nous n'avons pas connaissance des conditions de diffusion de cette ressource.

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, entre eux et par rapport à notre nouveau dictionnaire, pour les formes à trait d'union². Les deux premières colonnes donnent le nombre de lemmes et de formes dans chaque dictionnaire. Morphalou, avec 5608 formes, est le plus important, tandis que le Lefff est le plus petit. La colonne suivante (t. l.) donne le ratio nombre de lemmes / nombre de formes, ce qu'on peut appeler « taux de lemmatisation ». On voit que 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 ; l'écart du taux de lemmatisation entre ces deux dictionnaires s'explique par le fait que Multext contient 31 verbes (1207 formes) quand le Lefff n'en contient que 5 ; verbes mis à part, le taux de lemmatisation est de 1,76 pour les 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). Ainsi, par exemple, 29 % des formes du DELA se trouvent dans le

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. Par exemple, le Lefff contient six occurrences de la forme *sous-titre*, une comme nom, cinq comme formes du verbe *sous-titrer* ; ces six occurrences comptent pour une forme, et deux lemmes, dans notre tableau. 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).

Lefff. 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 dictionnaires. Le DM-C intègre de 61 à 90 % des formes des cinq dictionnaires que nous avons observés. Pour des raisons qui apparaîtront 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 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 : erreurs (ex. *hors-doeuvre*), segmentation différente (ex. *bras-le-corps*, cf. §2) ou expression d'un rapport (ex. *qualité-prix*).

	lemmes	formes	t. l.	dela	lefff	lexique	mph	multext	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) et M. Mathieu-Colas (1988) ont montré que des dictionnaires imprimés variaient sensiblement sur ce point ; cette variation se retrouve dans les dictionnaires électroniques. Les deux exemples donnés dans les premières lignes du tableau 2 parlent d'eux-mêmes. Ces variations sont en partie responsables des faibles taux de recouvrement relevés dans le tableau 1.

On sait aussi par ailleurs que les formes du **singulier** ou du **pluriel** de certains noms composés sont imprécises ou font débat. 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 : aucun dictionnaire ne connaît le pluriel *mots-valises*, le DELA ignore curieusement le singulier *année-lumière* et le Lefff contient des formes de pluriel qui paraissent fautives (il y a dans le Lefff une tendance à former des pluriels par simple ajout de *s* : *arc-doubleaus*, *extrême-onctions*, *sang-froids*...). 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 : comme les composés sont souvent cités comme sous-entrées du terme de base dans le TLF, l'information sur la flexion n'est pas signalée de façon formalisée et n'a donc pas pu être extraite dans Morphalou (cela explique aussi l'absence de lemmatisation déjà évoquée).

Un troisième axe de variation, moins souvent évoqué car plus spécifique aux dictionnaires pour le TAL, concerne la **segmentation** des unités lexicales. Le tableau 2 montre des variations autour des expressions (*de*) *plain-pied*, (*en*) *stand-by* et à *bras-le-corps*. 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.

UN DICTIONNAIRE ET UNE GRAMMAIRE DE COMPOSES FRANÇAIS

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

TABLE 2. Exemples de disparités entre dictionnaires.

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 lemmes avec *ex-*, 22 lemmes avec *co-*. Morphalou contient 123 formes commençant par *anti-* dont seulement 11 se retrouvent dans un autre dictionnaire. Morphalou est dérivé du TLF, or le TLF donne des composés de deux façons : directement comme mot vedette (p.ex. *presse-citron*), ou à l'intérieur des articles (p.ex. *pubo-fémoral* et *pubo-fémorien* figurent dans l'article *fémoral*). En particulier, le TLF a dans sa nomenclature des préfixes, suffixes, éléments de composition et autres « éléments formants », items pour lesquelles les articles donnent des listes de composés (p.ex. pour l'entrée *anti-* le TLF donne une longue liste de composés de fréquence ≥ 2 dans le corpus du TLF ; *presse-agrumes* figure sous l'entrée *presse-*, etc.). Morphalou, dans sa version 2, donne pour chaque mot l'indication de l'entrée d'origine, soit Morphalou 1, soit le TLF. Dans le second cas, qui s'applique à tous les mots contenant un trait d'union, on peut savoir si le mot relève d'un des trois cas de figure listés dans le tableau TABLE 3. On y voit que 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.

	nbre de mots	%
entrée vedette	2227	39,7
entrée secondaire sous un « élément formant » ³	2360	42,1
entrée secondaire sous un autre type de mot vedette	1021	18,2
TOTAL	5608	100

TABLE 3. Origine des 5608 mots à trait d'union de Morphalou dans le TLF.

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. Le premier est que, s'il sera aisé de reconnaître par exemple dans *assurance-chômage* un composé sur le schéma N-N, toute séquence N-N dans un texte ne sera pas nécessairement un composé. Dans un article sur la syntaxe du trait d'union, M. Mathieu-Colas (1995b) relève que celui-ci rentre parfois dans des structures complexes qu'il caractérise comme de trois types : corrélations (ex. *mi-sérieux, mi-plaisant ; mi-sel, mi-sucre*), juxtapositions (ex. *l'opposition consonnes-voyelles, la relation mère-enfant, le rapport signal-bruit, le trajet domicile-travail*) ou surcompositions (ex. *un porte-filtre à café, un ex-premier ministre, un contre-coup d'état, une conférence de presse-marathon*). L'existence de ces constructions fait qu'on ne peut considérer systématiquement une séquence xxx-yyy (p.ex. *presse-marathon*) comme un composé.

Surcompositions et juxtapositions sont un indice de la nécessité d'une prise en charge syntaxique des composés à trait d'union. Les surcompositions suggèrent une composition post-lexicale : on identifie par exemple *coup d'état* ou *conférence de presse* comme des composés lexicalisés (i.e. des mots figurant dans le dictionnaire), puis on reconnaît une construction préfixe-N dans *contre-coup d'état* et une construction N-N dans *conférence de presse-marathon*. Les juxtapositions sont, d'après (Mathieu-Colas, 1995b), « souvent précédées d'un nom introducteur » et leur reconnaissance relève donc de la syntaxe du groupe nominal. Il ressort de l'existence de ces constructions qu'en l'absence d'un système suffisamment sophistiqué pour distinguer les juxtapositions des compositions, le schéma de composition N-N ne peut être traité avec une précision absolue.

Le second problème auquel on s'expose dans l'identification des composés à trait d'union est posé par les ambiguïtés lexicales. Il est facile de caractériser des schémas tels que A-N (ex. *double-tierce*) ou V-N (ex. *serre-écrou*), mais il est plus difficile de ne pas associer à tort ces modèles à des séquences qui n'en relèvent pas ; par exemple, *ferme-auberge* ne désigne ni une auberge qui aurait une certaine fermeté, ni un instrument pour fermer les auberges. La distinction est importante si on considère que la reconnaissance des composés est aussi reconnaissance d'une composition sémantique.

Les difficultés posées par la composition N-N et par les ambiguïtés lexicales font que 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 TABLE 4 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 : *auto-école, auto-tamponneuse, micro-cravate, moto-ball, photo-finish, radio-réveil, stylo-bille..., compte-chèques, compte-rendu, garde-noble, grippe-influenza, guide-interprète, peigne-battant, peigne-détacheur, pince-monseigneur, porte-fenêtre, 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 4, deux règles permettant de construire ce type de suites avec ces formes ambiguës. La grammaire reconnaît donc par exemple *télé-surveillance* par deux schémas : PFX-N ou N-N.

³ On entend par « élément formant » des vedettes catégorisées dans le TLF non seulement comme *élément formant*, mais aussi comme *élément de compos.*, *élément de loc.*, (*élément*) *préf.* ou (*élément*) *suff.*

<i>Schémas</i>	<i>Exemples</i>
Composés sur préposition. Avec les prépositions <i>à, après, avant, en, hors, outre, pour</i> ou <i>sans</i> .	<i>à-côté, avant-programme, en-but, sans-opinion</i>
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, dont 13 sont marqués comme pouvant s'associer à un verbe.	<i>quasi-contractuellement, turbo-train, extra-plat, artério-veineux, sous-traiter, pré-enregistrer</i>
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 ⁴ .	<i>savoir-faire, bien-être, bien-aimé, attrape-couillon, cache-col, lave-linge, porte-bébé</i>
Composés numéral – nom pluriel.	<i>quatre-chevaux, trois-pièces</i>
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 ⁵ .	<i>blanc-seing, court-bouillon, double-toit, morte-saison, rouge-gorge, vif-argent</i>
Composés <i>sud, nord, est</i> ou <i>ouest</i> – adjectif.	<i>ouest-allemand, sud-vietnamien</i>
Composés verbe ou nom – groupe prépositionnel (V-GP et N-GP).	<i>boit-sans-soif, tape-à-l'œil, abri-sous-roche, belle-d'onze-heures</i>

TABLE 4. Types de composés identifiés par la grammaire.

Le schéma le plus productif dans cet ensemble, relativement aux dictionnaires, est de loin la composition avec préfixe (cf. tableau 5). 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 chaîne : en présence d'un composé comme *Vaires-sur-Marne*, en corpus, il faut éviter de reconnaître *sur-Marne* comme un composé PFX-N.

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 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.

⁴ Etant donné l'ensemble des composés V-N de Morphalou, on retient les verbes qui servent de base à au moins trois composés de type V-N, V-PRO (*brûle-tout*) ou V-GN (*trompe-la-mort*), où l'expression à droite du V a valeur de COD. Morphalou contient 23 entrées mettant en vedette un verbe comme « élément formant », ces 23 verbes figurent dans l'ensemble utilisé dans notre grammaire.

⁵ 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*.

A titre d'exemple, 247 + 140 (60 %) des 648 composés sur le schéma N-N (ex. *mot-valise*) sont associés au modèle M_S_C_S ou à sa variante féminine :

$$M_S_C_S = \langle E \rangle / m+s \mid s \langle P \rangle s / m+p ;$$

Cette déclaration se lit « étant donné la forme de base, ajouter la chaîne vide ($\langle E \rangle$) et associer les traits m et s , ou ajouter s , reculer d'un mot ($\langle P \rangle$), ajouter s et associer les traits m et p ».

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 :

$$\begin{array}{l} M_S_0 = \langle E \rangle / m+s \quad \left| \quad s / m+p+Rec \quad \left| \quad \langle E \rangle / m+p+Opt \quad ; \right. \\ M_0_S = \langle E \rangle / m+s+Rec \quad \left| \quad s / m+s+Opt \quad \left| \quad s / m+p \quad ; \right. \\ M_0_S_0_S = \langle E \rangle / m+s+Rec \quad \left| \quad s / m+s+Opt \quad \left| \quad \langle E \rangle / m+p+Opt \quad \left| \quad s / m+p+Rec \quad ; \right. \right. \end{array}$$

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. L'assignation de ces modèles de flexion aux composés a été déterminée par la consultation de (Catach, 1991), du *Petit Robert* et du TLF. Le DM-C est à notre connaissance le premier dictionnaire pour le TAL à intégrer de façon systématique la flexion nouvelle des composés.

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 5. Fréquence des types de composés.

L'information sur les schémas de composition associée aux entrées pourra servir pour l'étude des procédés de composition. On peut aussi la croiser avec les codes de flexion pour repérer les noms composés dont la flexion est singulière, p.ex. les composés sur phrase sont invariables à l'exception de *on-dit*, qui admet un pluriel avec ou sans *s* ; quoique dérivé d'une séquence N-A, *nerf-férure* prend un *s* au premier mot au pluriel (*nerfs-férures*) ; *un cent-garde* ou *une mille-feuille* s'écrivent sans *s* au singulier, contrairement aux autres composés NUM-N, etc.

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 vise à tolérer 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 chaîne 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 reconnaître 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 *super-héros* ou *extra-utérin*, où la soudure ne paraît ni en usage ni souhaitable. Les variantes de type *contreexpertise* / *contre-expertise* sont déclarées en deux entrées avec lemme commun.

Sont déclarés avec le caractère `=` (trait d'union ou espace), 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*), ainsi que 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*, *pole=position*). S'y ajoutent certains noms A-N, et certains adverbes ou adjectifs (ex. *bien=pensant*, *dos=à=dos*, *ex=aequo*).

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, ou les manuels d'orthographe le recommande dans les composés à valeur métaphorique comme *queue-de-pie*. Il nous paraît cependant, d'une part que cette norme est loin d'être appliquée systématiquement (cf. p.ex. videgreniers.org), d'autre part qu'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*, déjà cité. 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 (*blanc-bleu*, *blanc-gris*, *rouge-feu*, *rouge-sang*, *vert-jaune*). 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. (1990). « Un système de dictionnaires électroniques pour les mots simples du français », *Langue française* 87, 11-22. Paris : Larousse.
- 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. (1988). « Variations graphiques des mots composés dans le *Petit Larousse* et le *Petit Robert* ». *Linguisticae Investigationes*, XII:2.
- 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 ». *Linguisticae 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>.
- TROUILLEUX F. (2012). "Le DM, a French Dictionary for NooJ". VUČKOVIĆ K., BEKAVAC B., SILBERZTEIN M. (eds). *Automatic Processing of Various Levels of Linguistic Phenomena : Selected Papers from the NooJ 2011 International Conference*. Cambridge Scholars Publishing.
- VÉRONIS J. (1998). *Multext-Lexicons. A set of Electronic Lexicons for European Languages*. ELRA Catalogue (<http://catalog.elra.info>), MULTTEXT Lexicons, ref.: ELRA-L0010.