Morfetik, ressource lexicale pour le TAL

Pierre-André Buvet, Emmanuel Cartier, Fabrice Issac, Yassine Madiouni, Michel Mathieu-Colas, Salah Mejri

(1) CNRS LDI UMR 7187 – Université Paris 13, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse
{pabuvet,ecartier,fissac,ymadiouni, mmathieu-colas,smejri}@ldi.univ-paris13.fr

Résumé

Le traitement automatique des langues exige un recensement lexical aussi rigoureux que possible. Dans ce but, nous avons développé un dictionnaire morphologique du français, conçu comme le point de départ d'un système modulaire (*Morfetik*) incluant un moteur de flexion, des interfaces de consultation et d'interrogation et des outils d'exploitation. Nous présentons dans cet article, après une brève description du dictionnaire de base (lexique des mots simples), quelques-uns des outils informatiques liés à cette ressource : un moteur de recherche des lemmes et des formes fléchies ; un moteur de flexion XML et MySQL ; des outils NLP permettant d'exploiter le dictionnaire ainsi généré ; nous présentons notamment un analyseur linguistique développé dans notre laboratoire. Nous comparons dans une dernière partie Morfetik avec d'autres ressources analogues du français : Morphalou, Lexique3 et le DELAF.

Abstract

Automatic language processing requires as rigorous a lexical inventory as possible. For this purpose, we have developed a morphological dictionary for French, conceived as the starting point of a modular system (*Morfetik*) which includes an inflection generator, user interfaces and operating tools. In this paper, we briefly describe the basic dictionary (lexicon of simple words) and detail some of the computing tools based on the dictionary. The computing tools built on this resource include: a lemma / inflected forms search engine; an XML and MySQL engine to build the inflected forms; the generated dictionary can then be used by various NLP Tools; in this article, we present the use of the dictionary in a linguistic analyser developed at the laboratory. Finally, we compare Morfetik to similar resources : Morphalou, Lexique3 and DELAF.

Mots-clés : dictionnaire morphologique du français, CMLF, analyse linguistique des textes

Keywords: French morphological dictionary, XML, CMLF, Linguistical analysis, Morfetik

1 Structuration des données linguistiques

La ressource lexicale visée par le projet Morfetik est un dictionnaire morphologique du français. Cette ressource est le résultat du travail d'une vingtaine d'années de collecte et de description, sous la direction de Michel Mathieu-Colas. Nous présentons ci-dessous les informations essentielles sur cette ressource. Pour une présentation linguistique détaillée, voir (Mathieu-Colas, 2009).

1.1 Sources d'information

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. Les principales sources consultées sont : 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. 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 ainsi été identifiés, répartis comme suit :

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.

1.2 Description des entrées : tables de lemmes et tables de flexions

La structure des tables étant différente selon les catégories morphosyntaxiques, nous avons mis en place 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 tables 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.

A titre d'exemple, nous présenterons ici les encodages retenus pour les verbes, de loin la partie du discours morphologiquement la plus complexe en français. 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.

Au total, 222 codes de flexion ont été définis, et la table des flexions comprend les champs suivants :

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

Table 1 : 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.

Il est intéressant de noter que certains modèles intègrent des variantes (elles sont séparées par des points-virgules dans les tables de codes). Il peut s'agir par exemple d'une forme isolée (ECLORE [100.1] : ind. prés. [3s] = *il éclôt* ou *il éclot*), ou 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...*).

Nous renvoyons à (Mathieu-Colas, 2009) pour une présentation exhaustive des tables liées à chaque partie du discours.

1.3 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. A terme, d'autres informations seront intégrées à cette base, qu'il s'agisse de la nomenclature (ajout de néologismes et de termes spécialisés) ou d'informations descriptives complémentaires.

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.

2 Exploitations informatiques de la ressource linguistique

La ressource linguistique brièvement présentée ci-dessus est exploitée informatiquement. Le schéma ci-dessous synthétise les différentes exploitations du système :

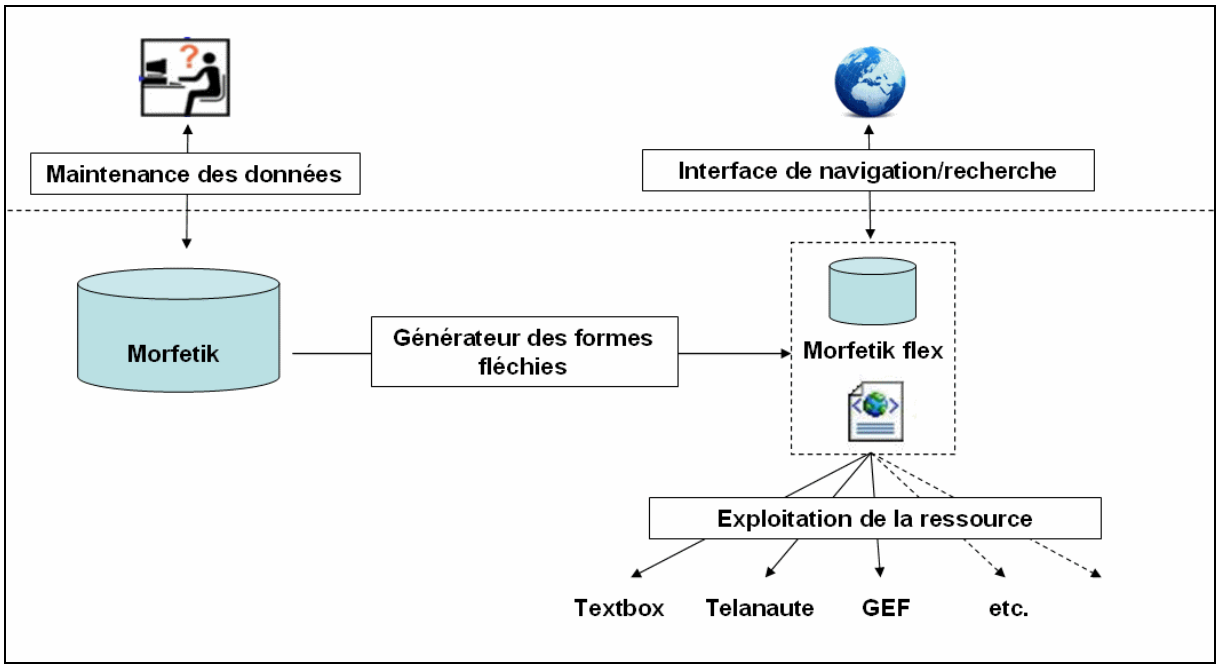

Figure 1 : architecture générale (Morfetik et ses outils)

2.1 Le moteur de flexion : Morfetik Engine

Morfetik Engine est la composante du système qui prend en charge l'interprétation de Morfetik Tables pour générer le lexique des formes fléchies, Morfetik Flex. Il comprend un moteur principal ainsi que des modules. Le moteur principal se charge de la collecte des informations dans les tables, de la création des formes fléchies à partir du lemme et du code de flexion approprié, de l'attribution des catégories grammaticales correspondantes (genre, nombre, etc.) et des autres métadonnées (par exemple rareté), et enfin il centralise l'exécution des modules. Ces modules sont de plusieurs ordres. On trouve tout d'abord les modules gérant les différents formats de sortie : actuellement, il existe un module pour générer une table mysql des formes, et un autre pour générer un fichier au format XML. Ensuite, chaque partie du discours nécessitant une flexion dispose d'un module spécifique permettant de gérer finement les particularités des tables qui lui correspondent, comme les différents types de codes pour la rareté, ainsi que les différents moyens de signaler les formes défectives ou les variantes. On trouve donc à l'heure actuelle un module [Nom], un module [Adjectif] et un module [Verbe]. Cette architecture modulaire permet de pouvoir assez facilement envisager la réutilisation du moteur pour d'autres langues ou d'autres ressources.

2.2 Morfetik Flex : l'ensemble des formes fléchies

Morfetik Flex est la ressource linguistique la plus directement utilisable par des applications informatiques puisqu'elle consiste en l'ensemble des formes fléchies de la ressource. A chaque forme sont associés un lemme ainsi que différentes informations de catégorie, de genre et de nombre (noms, adjectifs) ou de temps, de mode, de personne et de genre (verbes).

Cette ressource ayant pour vocation à être partagée, il était important de choisir un format d'échange respectueux des normes. C'est ainsi que nous avons notamment décidé de générer Morfetik Flex dans un format XML utilisant une DTD proche de celle proposée par le lexique Morphalou (cf. infra), cette DTD s'appuyant elle-même sur la norme LMF (Lexical Markup Framework).

2.3 Navigation et interrogation de Morfetik Flex

Un ensemble d'outils de navigation et de consultation de la ressource sont actuellement en finalisation de développement et seront mis à la disposition du public au premier semestre 2009 :

- outil de lemmatisation de formes : à partir d'une forme quelconque, l'outil renvoie l'ensemble des lemmes pouvant y correspondre; cette lemmatisation est également utilisée par les analyseurs développés au laboratoire (cf. section 2.5.) ; par exemple, si nous tapons « marche », en prenant l'option « étendre la recherche en enlevant/ajoutant les accents », l'interface nous renverra le tableau suivant, qui exhibe les différentes analyses possibles de la forme :

Lemme	Categorie	Temps	Nombre	Genre	Personne
marcher	Vrb	Ind_pr	S		1
marcher	Vrb	Ind_pr	S		3
marcher	Vrb	Sub_pr	S		1
marcher	Vrb	Sub_pr	S		3
marcher	Vrb	Imp_pr	S		2
marcher	Vrb	Pp	S	M	
marche	Nom	NULL	S	F	NULL
marché	Nom	NULL	S	M	NULL

Figure 2 : résultat d'analyse pour la forme « marché »

- outil de génération des flexions : à partir d'une zone permettant de saisir une forme, l'outil fournit les différentes lemmes possibles, puis les différentes formes liées à chaque lemme; par exemple, si nous donnons au générateur la forme « être », il nous renvoie à la fois la forme nominale et la forme verbale. Ensuite, l'utilisateur peut cliquer sur l'une des formes pour obtenir les formes correspondantes :

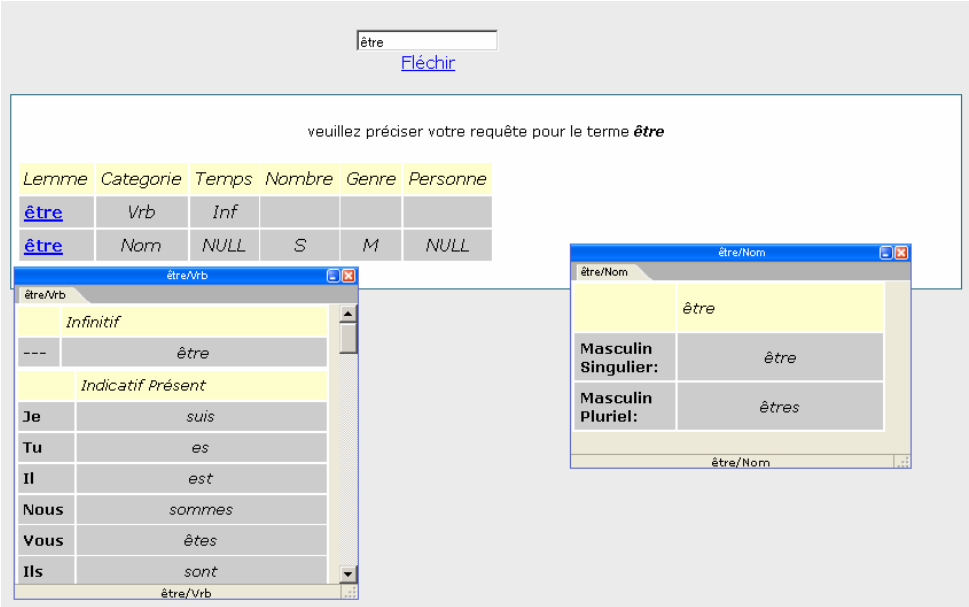

Figure 3 : résultat de la recherche pour être(2)

- moteur de recherche sur les formes et/ou les flexions : enfin, un moteur de recherche permet de lancer des recherches classiques sur l'ensemble des champs disponibles dans la base, par exemple pour rechercher tous les noms qui finissent par « on » ou encore les noms masculins qui sont aussi des verbes, etc. L'idée est de proposer aux chercheurs un moteur de recherche permettant d'effectuer des recherches complexes sur la base lexicographique.

2.4 Interface de maintenance

Actuellement, la maintenance, comme présentée dans l'architecture, a lieu sur les tables de lemmes et de flexions, et non pas sur les résultats flexionnels. De ce fait, l'interface de maintenance concerne les deux séries de tables propres à chaque partie du discours. Actuellement, la base est hébergée sous mysql, avec des connecteurs permettant de travailler à partir de masque ACCESS, ou bien directement via les masques proposés par phpmyadmin.

2.5 Exploitation dans d'autres outils

Morfetik Flex est l'outil principal pouvant être exploité de différentes façons. Dans le laboratoire, trois applications principales ont actuellement été développées : 1/ intégration de la ressource dans un outil d'analyse linguistique des textes (TextBox) ; 2/ intégration dans une chaîne de repérage des néologismes d'une langue (Telanaut) ; 3/ intégration dans un générateur automatique d'exercices de français dans une plateforme de e-learning (GEF). Nous présentons dans la section suivante l'exploitation de Morfetik Flex dans TextBox.

2.5.1 Morfetik comme ressource pour l'analyse morphologique des textes : TextBox

La ressource Morfetik est actuellement utilisée par un analyseur linguistique des textes développé au laboratoire, appelé TextBox (Cartier, 2007 ; Cartier, 2008). Cet analyseur permet, à l'aide de ressources linguistiques externes, d'effectuer l'analyse de textes numériques en plusieurs étapes : normalisation des textes en XML, segmentation en phrases

et en mots, analyse morphologique, puis analyse syntactico-sémantique. Chacune des étapes recourt à des ressources linguistiques externes : définition des « mots » et des « phrases » sous forme de listes d'expressions régulières pour l'étape de segmentation, dictionnaire morpho-syntaxique des formes pour la seconde étape. Dans ce cadre, TextBox a recours, pour le français, au dictionnaire Morfetik. TextBox annote, à chaque étape de l'analyse, le texte de nouvelles informations, sous forme d'éléments XML avec d'éventuels attributs. Voici, à titre d'exemple, le résultat de l'analyse de la séquence *Lorsque Renault est entré dans le capital de Nissan :*

```
- <p>
  - <token typo="tc" sign="word">
      <morph cat="conjs" lemma="lorsque" />
      Lorsque
    </token>
    <token sem="FrOrg" sign="word">Renault</token>
  - <token typo="lc" sign="word">
      <morph cat="noun" number="singular" lemma="est" gender="masculine" />
      <morph tense="ind" cat="verb" number="singular" lemma="être" />
      est
    </token>
  - <token typo="lc" sign="word">
      <morph tense="ppast" cat="verb" number="singular" lemma="entrer" gender="masculine" />
      entré
    </token>
  - <token typo="lc" sign="word">
      <morph cat="prep" lemma="dans" />
      dans
    </token>
  - <token typo="lc" sign="word">
      <morph cat="det" number="singular" lemma="le" gender="masculine" />
      <morph cat="pronoun" number="singular" lemma="le" gender="masculine" />
      le
    </token>
  - <token typo="lc" sign="word">
      <morph cat="adj" number="singular" lemma="capital" gender="masculine" />
      <morph cat="noun" number="singular" lemma="capital" gender="masculine" />
      capital
    </token>
  - <token typo="lc" sign="word">
      <morph cat="det" lemma="de" />
      <morph cat="prep" lemma="de" />
      de
    </token>
    <token sem="FrOrg" sign="word">Nissan</token>
    <token type="other" sign="punct">,</token>
```


Figure 4 : exemple d'analyse par TextBox-Morfetik

Cet exemple explicite, pour chaque « token » (c'est-à-dire chaque « mot » typographique reconnu dans la phase de segmentation), les différentes analyses morphologiques, hors contexte, contenues dans le dictionnaire morphologique ; chaque analyse débute par un élément « morph » doté d'attributs correspondant aux propriétés morpho-syntaxiques attachées à chaque forme : c'est ainsi que *lorsque* est une conjonction de subordination (cat=conjs), et que son lemme est 'lorsque' ; par contre, la forme *est* dispose de deux analyses possibles : comme nom et comme verbe ; de même pour *capital*, qui peut être adjectif ou nom.

Actuellement, une interface permet de visualiser, à partir des analyses morphologiques de Textbox et de Morfetik, les analyses possibles pour chaque mot, ainsi que les mots « inconnus » du dictionnaire (voir figure 7) : on constate évidemment que les noms propres seront inconnus, ainsi que de nombreux composés (en cours d'introduction dans Morfetik) et mots préfixés.

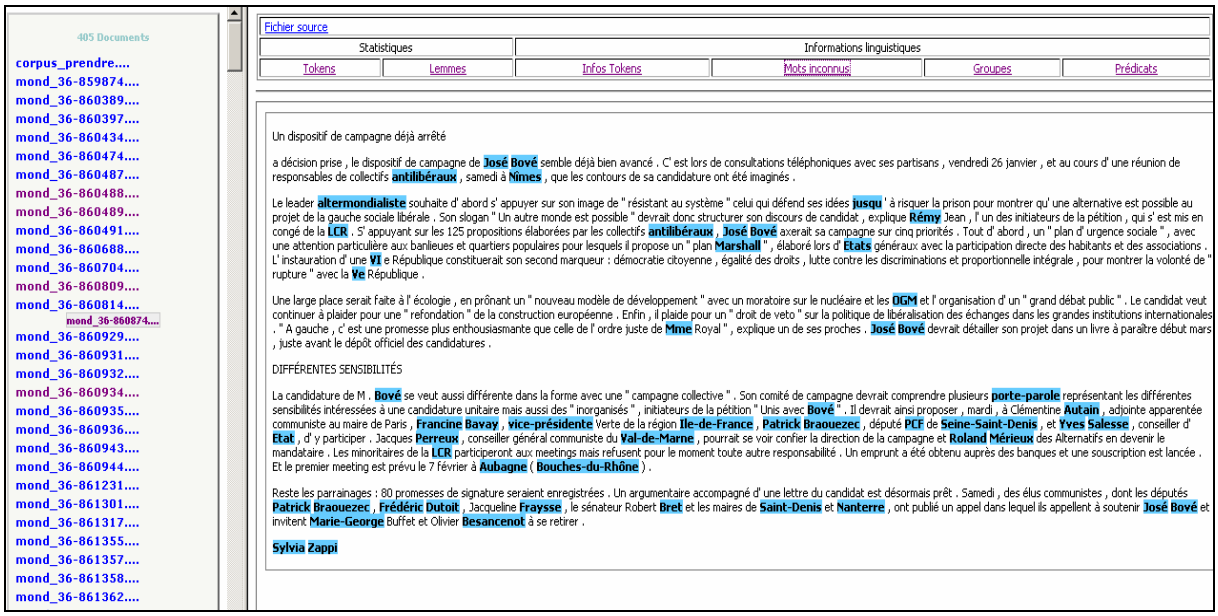

Figure 5 : Interface de TextBox pour debugger Morfetik

3 Comparaison avec d'autres ressources linguistiques du français

La présente section effectue une brève comparaison entre Morfetik et d'autres ressources linguistiques analogues. On trouvera en fin de section un tableau récapitulatif sur les différentes propriétés des ressources citées ici.

3.1 Morphalou

Morphalou est une ressource comparable à Morfetik Flex. Ce lexique comprend des formes complexes, ce qui n'est pas encore le cas de Morfetik. Sa couverture est cependant moindre d'environ 220 000 termes. Les approches des deux ressources concernant le genre des noms sont similaires (séparation des genres). Morphalou maintient cependant un lien entre les noms masculins et féminins, lien qui n'est pas encore mis en place dans Morfetik. Les indications de rareté présentes dans Morfetik permettent, quant à elles, de désigner de nombreuses formes peu courantes mais attestées comme rares, plutôt que de les considérer comme défectives. D'autre part, comme les deux ressources partagent des DTD très semblables, on peut dire qu'elles répondent à des scénarios d'utilisation similaires. De nombreuses différences fines existent toutefois, bien qu'elles dépassent le cadre de cet article. Notons seulement une différence importance dans la méthode de constitution du lexique, puisque Morphalou fonctionne par extraction des formes fléchies du TLF, puis correction des erreurs générées, à la main ou automatiquement. En revanche, Morfetik Flex est généré de manière systématique à partir de ses ressources Morfetik Tables, sans intervention manuelle. Quand un bug apparaît, les corrections sont apportées aux données sources ou bien au moteur lui-même, jamais au lexique.

3.2 Lexique 3

Lexique 3 est également une base lexicale regroupant des formes fléchies avec de nombreuses informations les concernant. Si la quantité de formes présentes est inférieure à celle de Morfetik, Lexique se distingue en revanche par la présence de nombreux autres renseignements (phonologie et fréquences entre autres). De plus, il dispose d'un éventail intéressant d'outils connexes comme la recherche de cooccurrences sur corpus ou la conversion vers le formalisme de Morphalou.

Nom	Mots simples	Mots composés	Indic. De Fréquence	Mode de création/origine	Format	Licence	Outils
Morfetik	758 152	Env. 140 000 (dont 120000 noms) Travail en cours	Marqueur de rareté	Automatique à partir de tables de lemmes et de flexions	XML-LMF	En cours de définition	Outil de consultation en ligne
Morphalou	524 725	Très peu (moins de 1%)	Non	Ad hoc à partir du TLF	XML-LMF	Licence spécifique autorisant la réutilisation	Non
Lexique3	150000	Très peu (moins de 1%)	Oui	Ad hoc	Tabulaire	Apparenté à la GPL	Scripts
DELAF	683 824	108 436	Non	Automatiquement à partir du DELAS	XML	LGPLLR (apparenté à GPL)	Non

3.3 DELAF

Le dictionnaire DELAF (flexions dérivées du DELAS) est sans doute, pour les formes fléchies du français, la ressource lexicale de large couverture la plus ancienne. Elle a spécifiquement été développée pour être utilisée dans le cadre d'une application informatique. La version XML est disponible sur Internet (http://infolingu.univ-mlv.fr/). Bien que Morfetik ait pris en compte, dans son élaboration, les acquis du DELAS, les différences concernent à la fois l'établissement de la nomenclature (sources utilisées) et la structuration des données (représentation des noms, marquage des formes verbales rares, etc.).

Conclusion

Morfetik est actuellement la ressource morphologique du français la plus exhaustive. Cette ressource dispose de différents modules qui en permettent la maintenance et l'exploitation : tout d'abord la ressource elle-même est gérée au laboratoire par le biais d'une base de données MySql qui est accessible via une interface web ou via des masques MS Access. Ensuite, cette ressource dispose d'un générateur de formes qui permet de disposer de l'ensemble des formes fléchies au format XML (CMLF). Pour les chercheurs, un premier accès est prochainement prévu pour visualiser et naviguer dans la base. Cette ressource XML peut ensuite être exploitée par des outils de TAL, et nous avons présenté dans cet article un analyseur qui intègre cette ressource. Morfetik sera consultable dès septembre 2009 à l'adresse suivante : http://extranet-ldi.univ-paris13.fr/Morf. On pourra également consulter à l'adresse http://extranet-ldi.univ-paris13.fr/ecartier/textbox une interface présentant un certain nombre d'analyses textuelles automatiques effectuées par TextBox à partir de Morfetik.

Références

BUVET PIERRE-ANDRE, 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.

CARTIER EMMANUEL (2007), « TextBox, a Written Corpus Tool for Linguistic Analysis », Web As Corpus 2007, Louvain-la-Neuve, sept. 2007.

CARTIER EMMANUEL (2008), Intégration des prédicats verbaux dans l'analyseur sémantique TextBox : l'exemple des verbes de cognition », Verbum, numéro spécial sur les prédicats verbaux, Paris.

COURTOIS BLANDINE (1990). « Un système de dictionnaires électroniques pour les mots simples du français », in COURTOIS Blandine, SILBERZTEIN Max (eds), *Dictionnaires électroniques du français, Langue française*, n°87, Paris, Larousse, pp. 11-22.

COURTOIS BLANDINE, SILBERZTEIN MAX (1990). *Dictionnaires électroniques du français, Langue française* n° 87, Paris, Larousse.

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

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

MATHIEU-COLAS MICHEL (2009, à paraître). *Morfetik*, une ressource lexicale pour le TAL, Cahiers de Lexicologie, Paris.

NEW B. (2006), *Lexique 3 : Une nouvelle base de données lexicales*. Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006), avril 2006, Louvain, Belgique.

ROMARY L., SALMON-ALT S., FRANCOPOULO G. (2004). *Standards going concrete : from LMF to Morphalou.*