

Titre du projet : **LEGERe : Un LExique Génératif de Référence pour le français**

Acronyme ou titre court (maximum 12 caractères) : **LEGERe**

Durée du projet en nombre de mois (entre 12 et 36) : 24 mois
Date de démarrage prévue : Juin 2008

Nom et prénom du coordonnateur du projet : Fiammetta Namer
Statut et établissement : PR, Linguistique, UFR Sciences du Langage, Université Nancy2
Laboratoire d'appartenance (code unité et intitulé) : UMR 7118 ATILF (Analyse et Traitement Informatique de la Langue Française)
Téléphone : 03 83 96 71 93 adresse email : fiammetta.namer@univ-nancy2.fr
Etablissement maître d'ouvrage du projet : UMR ATILF

Axe du programme scientifique concerné : AXE 2, Langues, Textes et Documents
Ce projet fait-il suite à une pré-opération : NON

3 à 5 mots clés : Lexique du français, Morphologie lexématique, Lexique Génératif, Traitement Automatique des Langues, Information Lexicographique

Tableau des partenaires du projet :

| | Nom resp. sc. | Prénom | Discipline | Laboratoire Nom et numéro d'unité | Etablissement de rattachement |
|---------|---------------|-----------|---|-----------------------------------|---|
| Coord. | NAMER | Fiammetta | morphologie, TAL | ATILF, UMR 7118 | Université Nancy2 |
| Part. 1 | BOUILLON | Pierrette | sémantique lexicale, lexique génératif, | ETI/TIM/ISSCO | Université de Genève, Ecole de Traduction et d'Interprétation, Traitement d'information multilingue |

Résumé du projet (*susceptible d'être publié sur le site MSH, maximum 1/2 page*)

1. Contexte scientifique et objectifs du projet
2. Description du projet, méthodologie
3. Résultats attendus

Le projet LEGERe a pour objectif la conception et constitution d'un lexique sémantique du français en vue de son utilisation en TALN, et de sa mise à disposition auprès de la communauté, via la plateforme du CNRTL. La conception de ce lexique repose sur deux types d'informations complémentaires acquises (semi-)automatiques ; celles issues des règles de construction de lexèmes, par l'utilisation de l'analyseur morphologique DériF, et celles issues de l'exploitation du corpus lexicographique du TLF. Parmi les résultats attendus, seront traités par la morphologie : les adjectifs en -able (lavable), les verbes dénominaux (déneiger, emprisonner) et déadjectivaux (banaliser, électrifier) et les noms de procès (lavage, construction, gonflement) ; l'acquisition à partir du TLF se focalisera sur les noms d'instruments non construits (balai), les noms composés N prep N (grain de blé, boîte à gants).

Les résultats, qui prévoient la création en deux ans d'un lexique de 42 000 entrées annotées sémantiquement, s'inscrivent dans le cadre formel du Lexique Génératif (LG), et relèvent du courant lexématique de la morphologie.

Les partenaires du projet réunissent donc des compétences complémentaires indispensables à la réalisation de ces objectifs : P Bouillon (LG, sémantique lexicale), I. Chovanová (composition nominale), Georgette Dal (morphologie), E. Jacquy (LG, TLF), F Namer (Morphologie, DériF).

Si possible, abstract en anglais

LEGERe is a project whose twofold aim is the design and the realization of a semantic lexicon for French which can be reused in NLP applications, and which will be put at researchers' disposal, through the CNRTL platform. This lexicon's design relies upon two types of complementary features, both acquired in a (semi)-automatic way : (1) information coming from lexeme formation rules, by means of a morphology parser called DériF, (2) information coming from the exploitation of the TLF dictionary lexicographic content. Among expected results, morphology will take care of the following lexical data : -able ending adjectives (lavable – washable), prefixed denominal verbs (déneiger – remove snow, emprisonner – put smne in jail), deadjectival verbs, be they prefixed (amaigrir – make (smne) slimmer) suffixed (banaliser – make smth trite, électrifier- electrify) or converted (jaunir – (make smth) turn to yellow). As far as TLF acquisition methods, they will permit to encode simple instrument nouns (balai – broom), and « N prep N » compounds (farine de blé : wheat flour, boîte à gants : glove compartment).

As a main result, we expect the creation in two years of circa 42,000 semantically tagged lexical entries. The annotation system comes within the scope of the Generative Lexicon (GL) formal framework ; as far as morphology values are concerned, they reflect lexematic formation theoretical assumptions.

LEGERe partnership gathers complementary skills that are necessary to realize these above-mentioned objectives: P Bouillon (GL, lexical semantics), I. Chovanová (compoundings), Georgette Dal (morphology), E. Jacquy (GL, TLF), F Namer (Morphology, DériF).

B. PROJET SCIENTIFIQUE

1. Introduction et présentation des objectifs généraux du projet, situation dans le programme scientifique de la MSH Lorraine (1 page)

Le projet LEGERe a pour objectif la conception et constitution d'un lexique sémantique du français en vue de son utilisation en TALN, et sa mise à disposition auprès de la communauté, via la plateforme du CNRTL. La conception de ce lexique repose sur deux types d'informations complémentaires acquises (semi-)automatiquement ; celles issues des règles de construction de lexèmes, par l'utilisation de l'analyseur morphologique DériF, et celles issues de l'exploitation du corpus lexicographique du TLF. Parmi les résultats attendus, seront traités par la morphologie : les adjectifs en *-able* (*lavable*), les verbes dénominaux (*déneiger*, *emprisonner*) et déadjectivaux (*banaliser*, *électrifier*) et les noms de procès (*lavage*, *construction*, *gonflement*) ; l'acquisition à partir du TLF se focalisera sur les noms d'instruments non construits (*balai*), les noms composés *N prep N* (*grain de blé*, *boîte à gants*). Les résultats s'inscrivent dans le cadre formel du Lexique Génératif (LG) (Bouillon, 1997, Pustejovsky, 1995), et relèvent du courant lexématique de la morphologie. Les partenaires du projet réunissent donc des compétences complémentaires indispensables à la réalisation de ces objectifs : P Bouillon (LG), I. Chovanová (composition nominale), Georgette Dal (morphologie), E. Jacquy (LG, TLF), F. Namer (Morphologie, DériF).

Relativement au programme scientifique de la MSH Lorraine, le projet LEGERe s'inscrit dans le cadre de l'axe 2 de ce programme, "Langues, Textes et Documents". Les traitements sémantiques de corpus écrits, qui s'intègrent dans la problématique de l'exploitation des corpus peuvent avoir des vocations diverses selon les communautés : constitution de ressources de grande taille annotées sémantiquement, analyse de phénomènes sémantiques à grande échelle, et dans le domaine du Traitement Automatique des Langues (TAL), recherche d'information, classification de documents, fouille de textes, Web sémantique, etc. Mais quelle que soit leur vocation la plupart des traitements sémantiques de corpus de textes passent par l'utilisation de lexiques sémantiques réutilisables, de grande taille, dans un format standard, et extensibles. Or, bien que de nombreux travaux aillent dans ce sens, une telle ressource n'existe pas encore pour le français. De ce point de vue les objectifs du projet LEGERe répondent doublement à la thématique "exploitation et constitution de corpus" affichée par l'axe 2. L'ambition de produire une ressource lexicale sémantique pour le français y répond en effet de par sa nature mais aussi de par l'approche méthodologique choisie dans le cadre du projet.

Tout d'abord, l'existence d'une telle ressource pour le français rend possible notamment les traitements d'annotation sémantique de corpus textuels à grande échelle et donc tout un ensemble de travaux sur le français, fondés sur l'existence de données textuelles de grande taille et annotées sémantiquement. Sur ce point de plus, la ressource produite sera normalisée en partant du format de représentation du lexique génératif, partagé au sein du projet. A partir de cet unique format, nous collaborerons avec le CNRTL, structure adossée à l'ATILF, premièrement afin de normaliser le lexique produit selon les recommandations du format LMF, et deuxièmement, afin de mettre en oeuvre sa mise à disposition concrète auprès de la communauté scientifique, lorraine, nationale et internationale.

Ensuite, la méthodologie de constitution de la ressource lexicale étant fondée sur la consolidation et la collaboration de différentes techniques d'acquisition (fondations morphologique, courant lexématique, et exploitation d'un corpus lexicographique), elle permet d'envisager la construction d'un lexique cohérent (homogénéité de l'approche morphologique), d'un bon degré d'exhaustivité (appui sur le dictionnaire de référence du Trésor de la Langue Française informatisé) et suffisamment formalisé et robuste pour être utilisable dans le domaine du Traitement Automatique des Langues (appui sur le modèle théorique du Lexique Génératif et sur nos travaux antérieurs ayant fourni une extension à ce modèle pour y représenter les informations de nature morphologique, Jacquy & Namer 2007).

2. Description du projet scientifique (8 pages)

2.1. Objectifs

L'objectif du projet est la conception, puis la réalisation en deux ans, d'un lexique sémantique du Français de taille réelle, comportant 42 034¹entrées, et produit (semi)-automatiquement au moyen d'une application déjà existante : le programme DériF (Namer, 2002, Namer, 2003, Namer, 2005). Ce programme analyse automatiquement les mots construits morphologiquement en leur associant un ensemble de traits morphologiques, catégoriels et sémantiques. Ce programme ne traite qu'une partie du lexique, certes importante (il est à noter qu'à ce jour, DériF réalise l'analyse morphologique de 37,7 % des quelques 92 000 entrées de la nomenclature du TLFi). La couverture actuelle de DériF en matière de reconnaissance de règles de construction de lexèmes (RCL) inclut :

1. les RCL formatrices de noms déverbaux en *-ion, -age, -ment, -eur, -oir*, d'adjectifs déverbaux en *-able, -if*
2. les RCL formatrices de noms désadjectivaux de propriété en *-ité*
3. les RCL formant des verbes désadjectivaux et/ou dénominiaux par suffixation en *-iser, -ifier*, par préfixation en *dé-, é-, en-*, et par conversion
4. les RCL formatrices d'adjectifs dénominiaux en *-eux, -al, -ique, -ien, -aire, ...*
5. les RCL formatrices d'adjectifs en *anti-, in-, super-, hyper-, sous-, ...*
6. les RCL de composition néoclassique formatrices de noms et d'adjectifs dits savants.

DériF étant conçu suivant une approche récursive, il analyse par étapes successives tout lexème dont la construction est le produit de l'application de plusieurs des règles mentionnées ci-dessus. Les détails concernant DériF et les résultats attendus sont donnés au paragraphe (3.2).

Pour compléter la couverture, nous combinons l'application de DériF à l'utilisation d'indices lexicaux et structuraux, appris et réutilisés dans le cadre de l'exploitation du corpus lexicographique du TLFi. Cette méthode va en priorité s'intéresser à deux familles lexicales ignorées par DériF :

- les noms simples dont la définition permet d'identifier leur appartenance ontologique (entité animé : *chien*, végétale : *herbe*), ou que des marqueurs permettent de caractériser comme des outils (*balai, scie*).
- les composés pluri-lexicaux instanciant la séquence catégorielle "nom préposition nom" (*grain de blé, moule à gaufres*), dont l'analyse repose sur la caractérisation sémantique de la relation entre les deux noms en présence.

De plus, l'utilisation des définitions du TLF va permettre de compléter les informations acquises par l'analyseur morphologique. Les détails concernant l'acquisition d'informations à partir des corpus définitoires du TLF, et les résultats attendus, sont donnés au paragraphe (3.3).

Notre préoccupation principale, dans la production de ce lexique, est sa cohérence : quelle que soient la catégorie et le type sémantique des éléments qui constituent à terme cette ressource, ceux-ci doivent être représentables au moyen du même faisceau de traits ; les informations acquises doivent être réutilisables par des applications de différentes natures ; des vérifications croisées doivent pouvoir être effectuées par les membres du projet, qui doivent par conséquent partager des critères communs de validation. Pour répondre à ce souci de cohérence interne et de visibilité externe, deux décisions clé ont été prises :

- le lexique généré est composé exclusivement de lexèmes, qui, dans le sens de (Matthews, 1991) et de (Fradin, 2003), constituent la classe ouverte des unités à sens référentiel et munies de l'une des catégories majeures : Nom, Adj, Verbe et certains Adverbes. L'une des raisons, d'ordre méthodologique, à ce choix, est que DériF (le principal fournisseur d'informations structurées pour le projet) n'opère que sur des lexèmes, puisqu'en français contemporain, seul un élément à sens référentiel peut être la base ou le résultat d'une règle de construction morphologique. La deuxième raison est

¹ Parmi les entrées de ce lexique, 26 117 seront produites à partir des analyses de DériF, les autres seront issues de l'exploitation du TLFi : 9257 constructions *N prep N* avec *prep=à/de* et, d'après des recherches préliminaires, 6660 noms identifiables par leur fonction

d'ordre plus théorique : le sous-ensemble complémentaire aux lexèmes dans le lexique est formé des grammèmes (déterminants, connecteurs, auxiliaires, conjonctions, prépositions...). Leur rôle est à l'interface entre syntaxe et sémantique : ils n'appartiennent pas au lexique en tant qu'objet d'étude ; en tout cas le TAL ne s'attend pas à les traiter ni à les utiliser comme les noms, les verbes ou les adjectifs.

- le format pivot de travail s'inspire du modèle proposé dans le cadre du Lexique génératif (Bouillon, 1997, Pustejovsky, 1995). Plusieurs raisons ont dicté ce choix. Tout d'abord, le LG a été conçu dans le but de rendre compte de façon dynamique et économique de la polysémie des langues : une entrée sert de point de départ aux variations lexicales sémantiquement apparentées. Cette conception est donc également parfaitement adaptée à la représentation de néologismes, et rend donc possible une extension du lexique, par rapport aux objectifs envisagés dans le cadre strict du projet. Une autre raison est la possibilité d'accéder aux informations lexicales à travers différents angles de lecture. Le détail du format sera expliqué au paragraphe (3.1).

2.2. Déroulement du projet

Notre projet comporte deux phases :

Phase1 : Etudes, traitements et validations parallèles,

Phase2 : Etudes, traitements et validations complémentaires.

La **Phase1** sera réalisée lors de la première année du projet. La **Phase2** occupera une large partie de la deuxième année.

2.2.1. Première année

Lors de la **Phase1**, l'acquisition d'information par DériF et par l'exploitation du TLF se feront en parallèle, par la réalisation de deux tâches, sur des données différentes :

Tâche1 : acquisition par la morphologie (DériF) : verbes dénominaux et désadjectivaux, adjectifs déverbaux et dénominaux, noms déverbaux. Nous nous focalisons prioritairement sur les constructions mentionnées sous 1 - 4 au paragraphe (2.1).

Tâche2 : acquisition par le TLF : composés "Nom prep Nom", noms morphologiquement non construits.

La tâche1 consiste en la succession de plusieurs étapes faisant intervenir trois types de compétences : morphologie (construction du lexique du français) sous la responsabilité de G. Dal, TAL (analyse morphologique) sous la responsabilité de F. Namer et sémantique lexicale (validation du format de sortie et des traits codés) sous la responsabilité de P. Bouillon.

La tâche 2 se décompose en deux étapes principales, séquentielles dans un premier temps, mais qui donneront lieu ensuite à des allers-retours en fonction de la précision et du rappel des résultats obtenus. La première étape, réalisée sous la responsabilité de E. Jacquy, consistera à raffiner les méthodes d'exploitation du TLFi sur la base de travaux antérieurs. Cette première étape permettra de détecter des constructions de la forme *N prep N* (*chaussures de marche, saumons d'élevage, placard à balais*) ainsi que les noms simples (non construits) du français. La seconde étape se concentrera sur l'analyse linguistique des *N prep N*, sous la responsabilité d'I. Chovanovà, et celle des noms simples détectés (typage sémantique).

2.2.2. Deuxième année

La Tâche1 continuera son déroulement pendant la deuxième année du projet, et se focalisera sur l'analyse morphologique es lexèmes préfixés, convertis, et construits par composition néoclassique.

En parallèle commence la Phase2 du projet. Cette phase prévoit une collaboration entre les deux méthodes d'acquisition de manière à produire des résultats plus complets et précis. Ce raffinement des informations concernera les types de données suivantes :

- noms déverbaux d'action (et verbes de base) : *grondement, harnachement*
- noms déverbaux d'agent/instrument (et verbes de base) : *ajusteur, amortisseur*
- noms déverbaux d'instrument/lieu (et verbes de base) : *rasoir, dortoir*
- adjectifs déverbaux en *-able* (et verbes de base) : *lavable, circulaire*

L'acquisition d'informations à partir du TLF portera sur l'identification de : la structure argumentale du verbe (*circuler* se construit avec un locatif alors que *laver* est transitif direct), la nature agentive ou non de ses arguments (au sens de (Dowty, 1991), *affolement* a un argument de type proto-patient, à l'inverse *grondement* est proto-agentif), l'interprétation du nom le cas échéant (lieu ou instrument pour le nom en *-oir*, agent ou instrument, pour le nom en *-eur*).

Les résultats partiels de DériF seront ainsi complétés par les informations lexicographiques obtenues grâce à des indices lexicaux dans les définitions du TLF.

2.3. Membres du projet, et compétences

Membres de l'UMR ATILF, équipe Lexique

- Evelyne Jacquey (CR ATILF): sémantique lexicale, exploitation du TLF,
- Fiammetta Namer (PR Linguistique UMR ATILF – Université Nancy2) : morphologie, DériF, sémantique lexicale,
- Iveta Chovanová (doctorante à l'ATILF) : morphologie, sémantique des "Nom prep Nom",
- Georgette Dal (PR Linguistique, UMR STL-Université Lille3) : morphologie

Membre de l'université de Genève, Ecole de Traduction et d'Interprétation, Traitement d'information multilingue (ETI/TIM/ISSCO)

- Pierrette Bouillon (Professeure) : sémantique lexicale, lexique génératif, traitement automatique des langues

3. Description détaillée du projet

3.1. Le format pivot

Une entrée du lexique génératif est concevable comme le produit de quatre rôles sémantiques (FORMEL, CONSTITUTIF, AGENTIF, TÉLIQUE). Chaque rôle définit un prédicat qui relie entre eux des arguments logiques, et est caractérisé d'un point de vue événementiel. Les paramètres manipulés par le prédicat sont typés, ainsi que la structure qui englobe les quatre rôles appelée structure de qualias, désormais StrQua. Le typage de la StrQua (sous forme de Paradigme Lexical Conceptuel) organise le lexique en classes sémantiques.

La valeur de chacun des rôles mentionnés ci-dessus à une fonction déterminée pour définir le sens de l'objet M dénotée par l'entrée lexicale :

| ROLE | Fonction | Exemple (N) : couteau |
|-------------|--|--|
| FORMEL | place de M dans l'ontologie | x de type <i>artefact</i> |
| CONSTITUTIF | relations (partie-tout) entre M et ses constituants | Relation = Composant-assemblage , entre x (ie le couteau) et z (de type manche) et entre x et u (de type lame) |
| AGENTIF | conditions nécessaires (présupposées) à l'existence de M | Événement (accomplissement) : fabriquer, faisant intervenir un agent w et le résultat x |
| TÉLIQUE | décrit la finalité de M | Événement (accomplissement) : couper, faisant intervenir un agent w', l'instrument x et l'objet à couper : y |

Figure 1 : structure d'une entrée lexicale dans le LG

Que doit-on considérer comme entrée du lexique génératif ?

Face à des vocables qui correspondent à plusieurs emplois prédicatifs possibles, un certain nombre de décisions devront être prises, en fonction du type de polysémie et des mécanismes de dérivation sémantique offerts par le LG de manière à favoriser le caractère synthétique et économique du lexique produit. Ainsi :

- (a) Certains prédicats transitifs admettent une lecture de type accomplissement ou de type activité selon la définitude de leur complément : *Alfred répare les mobylettes* / *Alfred répare ta mobylette*.
- (b) De même, les verbes de mouvement à direction intrinsèque (avancer, reculer, bouger, marcher...) alternent entre une interprétation de type accomplissement et une interprétation de type activité selon que l'objet locatif désigne une destination (*Max marche vers le sommet de la falaise*) ou pas (*Max marche dans la forêt* / *Max marche*)
- (c) Certains prédicats alternent entre interprétation causative et résultative (*Le patron véreux a coulé l'entreprise*. *L'entreprise a coulé*)

Nous nous proposons de prévoir la représentation lexicale la plus complète, l'autre pouvant s'obtenir par filtrage des rôles et arguments pertinents, en vertu de de la sélection des paramètres de la structure argumentale par les prédicats valant les rôles de la structure de qualia : la lecture accomplissement sera donc privilégiée dans les cas (a, b) ci-dessus. En ce qui concerne (c), les spécifications formelles eront conformes aux hypothèses du LG (l'entrée sera sous-spécifiée, c'est à dire privée de tête événementielle).

3.2. Acquisition par analyse morphologique

Cette approche exploite les résultats fournis par l'analyseur morphologique DériF (Namer, 2002, Namer, 2003, Namer, 2005), qui sont reformatés de manière à être en conformité avec les notations du LG, suivant l'extension du modèle proposé dans (Jacquey and Namer, 2007, Namer and Jacquey, 2003).

Présentation de DériF

L'analyseur DériF a vu le jour dans le cadre du projet MorTAL. L'analyse par DériF d'un lexème catégorisé adapte les hypothèses théoriques relevant du courant lexématique de la morphologie (Anderson, 1992, Aronoff, 1994, Fradin, 2003). Basé sur l'application d'un système ordonné de règles, le mécanisme est récursif et permet la gestion des ambiguïtés, se réappliquant sur chaque (liste de) résultat obtenu précédemment. L'analyse morphologique d'un lexème construit sur une base elle-même construite est donc hiérarchisée. Le résultat est un triplet, la première partie retrace sous forme crochétée l'historique des étapes d'analyse, la seconde réunit les lexèmes résultats obtenus à chaque étape, et la troisième est constituée d'une formulation en langue naturelle de la relation morphologique liant l'input à son (ses) constituant(s) immédiat(s). Les néologismes sont analysés et pseudo-définis comme des mots régulièrement construits (ce qui est généralement le cas).

Résultats de DériF

L'analyse d'un lexème par DériF s'accompagne d'annotations, reflétant les contraintes de la règle morphologique appliquée, et pouvant porter sur le lexème construit L et sa base B (Namer, 2002, Namer, 2005). Par exemple, la préfixation en *dé-* construit des verbes sur base adjectivale (eg. saoul/dessaouler). L'adjectif est toujours **qualificatif** et décrit une propriété **transitoire**. Le verbe est soit **transitif causatif** (*le café salé dessaoule Max*) soit **intransitif inchoatif** (*Max dessaoule*).

```
dessaouler/VERBE ==> 2,ADJ/dé1/pre/VERBE+saoul/ADJ
                    "(Supprimer - Faire perdre) le caractère saoul"
saoul/ADJ:(_,temporaire,_,predicatif)
dessaouler/VERBE:  (causatif,transitif,[cause,theme])|
                    (resultatif,intransitif,[theme])
```

Figure 2 : Analyse du verbe dessaouler par DériF

Traduction des résultats dans le LG

En fonction des résultats produits par DériF, il est possible de générer automatiquement au plus les deux entrées au format LG (celles du lexème analysé L et de sa base B) reliées morphologiquement par la règle ayant produit le résultat.

Le niveau de spécification de chaque entrée dépend des informations produites lors de la phase d'analyse morphologique.

Par exemple, il est possible de prédire la StrQua suivante pour le verbe dessaouler (dont la Figure 3 donne une version simplifiée), qui indique la succession de situations suivante :

- 1) l'individu y est saoul (état initial, présumé dans le rôle AGENTIF)
- 2) l'agent x dessoule l'individu y ou y dessoule (accomplissement, dans le rôle AGENTIF)
- 3) y n'est plus saoul (état final, rôle FORMEL)

| ROLE | (V) dessaouler |
|-------------|--|
| FORMEL | non (saoul (e1 :état,y :individu)) |
| CONSTITUTIF | - |
| AGENTIF | Saoul(e0 :état,y) ET dessaouler_acte (e2 :accomplissement, x :agent, y) OU Saoul(e0 :état,y) ET dessaouler_acte (e2 :accomplissement, y) |

Figure 3 : Représentation simplifiée de dessaouler en LG à partir de l'analyse par DériF

En ce qui concerne l'adjectif saoul, en tant que base de dessouler, la seule information inférable à partir de la RCL par préfixation en dé- est qu'il s'agit d'une propriété représentée dans le cadre du LG sous la forme d'un état e affectant un individu y (cf. Figure 4 pour une représentation simplifiée).

| ROLE | (A) saoul |
|-------------|----------------------------|
| FORMEL | saoul(e :état,y :patient) |
| CONSTITUTIF | - |
| AGENTIF | - |

Figure 4 : Représentation simplifiée de saoul en LG à partir de l'analyse par DériF

3.3. Acquisition par l'exploitation du TLFi

Le TLFi fournit une description lexicographique exceptionnellement riche pour près de 90.000 lemmes du français (89.961 hors affixes divers) ; les informations lexicographiques présentes sont, outre les définitions et les indicateurs d'emploi, des exemples attestés extraits de la base textuelle FRANTEXT, des locutions figées ou semi-figées définies, des indications de domaine technique d'emploi corrélées à des informations spécifiques d'usage, etc. La version actuelle du TLFi pour la recherche est accessible dans un format XML propriétaire et documenté et elle dispose d'une version catégorisée et lemmatisée de l'ensemble des définitions (271.165) et des exemples (427.493). À partir d'une ressource d'une telle richesse sur le français du 19^{ème} et du 20^{ème} siècle, l'enjeu général réside dans la capacité à gérer l'hétérogénéité de l'information présente, tant sur le plan formel que sur le plan des contenus. Avec le projet LEGERE, cet enjeu général se verra enrichi par la nécessité de mettre en correspondance l'organisation lexicographique de l'information sémantique avec l'organisation de cette même information sémantique dans le modèle LG.

Dans la durée du projet, nous lancerons trois chantiers principaux, le premier devant être finalisé avant les deux suivants qui eux, peuvent être la plupart du temps réalisés en parallèle. (3) sera réalisée dans la phase 2 du projet.

4. procédures d'exploitation du corpus lexicographique issu du TLFi
5. application pour l'extraction des locutions et expressions figées de la forme *N prép N* en français, et le typage sémantique des noms simples,
6. application pour l'extraction d'informations sémantiques complémentaires à celles de DériF sur les lexèmes construits en *-age, -ment, -ion, -oir, -eur, -able*. Concernant les

lexèmes en *-age*, *-ment*, *-ion* et *-able*, les procédures d'acquisition seront spécialisées afin d'extraire des informations sémantiques sur la structure argumentale de ces noms et des verbes de base. Concernant les noms en *-eur*, *-oir*, les procédures d'acquisition permettront de déterminer leur type sémantique entre /instrument/ et /personne/ avec *-eur* et entre /instrument/ et /lieu/ avec *-oir*.

3.3.1. Acquisition d'informations sémantiques à partir du TLFi

L'exploitation du TLFi s'appuiera sur des travaux antérieurs menés par plusieurs membres de l'équipe Lexique avec le soutien de l'équipe des informaticiens au laboratoire. Ces travaux ont permis de mettre au point un des procédures qui pourront être assemblées dans le cadre de ce projet :

- (1) détermination de la fréquence d'apparition d'un lexème dans le corpus des définitions (étant donné un lexème L, quel est l'ensemble des autres lexèmes qu'il permet d'atteindre lorsqu'il apparaît en début de définitions) – cette première procédure consiste donc à traiter les lexèmes dans les définitions comme des indices lexicaux ;
- (2) extraction des informations complémentaires pertinentes pour les définitions en fonction de la structure XML du dictionnaire.

3.3.2. Détection et modélisation de lexèmes non construits

Une expérience préliminaire, effectuée sur la version XML catégorisée du TLF, montre par exemple que les définitions de ce dictionnaire sont suffisamment régulières pour permettre de détecter les substantifs ayant une fonction prototypique, c'est-à-dire une facette télique dans leur contenu sémantique. En recherchant, dans le corpus des définitions, des expressions comme « servir, permettre, destiné à/au/aux/de », on repère que 14% de substantifs du TLF ont un emploi télique facilement détectable. On donne quelques exemples ci-dessous.

ABRI : *Lieu servant à protéger*

AGRAFE : *Petit objet métallique servant à attacher deux ou plusieurs choses ensemble*

AGOGE : *Dans les mines, rigole servant à l'évacuation des eaux*

BALAI : *Ustensile de ménage servant au nettoyage*

Des définitions comme celles qui sont retranscrites ci-dessus fournissent plusieurs types d'informations : (1) ces quatre substantifs ont une facette télique du point de vue du modèle LG, (2) l'étiquetage catégoriel des définitions permet d'acquérir aussi une facette formelle (les substantifs ou groupes nominaux qui précèdent directement l'expression recherchée, ici « servir à »), (3) la nature des prédicats représentant la facette télique – substantif prédicatif, « évacuation », « nettoyage » – ou – verbe, « protéger », « attacher ». Ainsi à partir d'une première recherche simple, on obtient des résultats de la forme suivante :

ABRI : *Lieu_{FORMEL} servant à protéger_{TELIQUE}*

AGRAFE : *Petit objet métallique_{FORMEL} servant à attacher_{TELIQUE} deux ou plusieurs choses ensemble*

AGOGE : *Dans les mines, rigole_{FORMEL} servant à l'évacuation_{TELIQUE} des eaux*

BALAI : *Ustensile de ménage_{FORMEL} servant au nettoyage_{TELIQUE}*

A l'issue de la mise au point des requêtes à appliquer au TLF, leur lancement sera réalisé sous la forme de traitements de corpus exploitant la structure XML des données et leur étiquetage grammatical. La définition d'un corpus de validation et l'utilisation d'autres traitements plus classiques permettront ensuite d'évaluer notamment le rappel et la précision de chaque type de requête. Enfin, l'ensemble des prédicats acquis à partir du TLF sera traduit dans le format fourni par le modèle LG afin d'être intégrés dans le lexique produit.

3.3.3. Modélisation des locutions et expressions figées de la forme N prep N

Le TLFi contient différentes sortes d'objets lexicographiques. Les constructions particulières sont identifiables d'abord grâce à la balise XML <syntita n="d"> ; ensuite, pour ne garder que les constructions de la forme N prep N, on se sert d'un étiqueteur grammatical (TreeTagger). Cet ensemble de constructions sera analysé sur le plan linguistique en prenant appui sur les modélisations proposées pour l'italien par F Busa et M Johnston (Busa, 1997, Johnston and Busa, 1996), et pour l'anglais, le turc et le français par Ch Bassac et P. Bouillon (Bassac and

Bouillon, 2005, Bassac, 2006). Par exemple, ces auteurs ont mis en évidence des relations entre *N1* et *N2* dans *N1 prep N2* : dans *chaussures de marche*, *marche* désigne la **fonction** des *chaussures*, alors que dans *saumons d'élevage*, *élevage* qui pourtant est aussi un nom déverbal désigne ici l'**origine** des *saumons*. Ces deux structures constituent donc des instances de deux modèles caractéristiques qui seront représentés de manière distincte dans le format pivot du lexique produit.

3.3.4. Structure argumentale des noms prédictifs en -age, -ment, -ion

Grâce à l'implémentation dans DERIF des règles de construction des lexèmes, les noms déverbaux en -age, -ment et -tion conduisent à un squelette sémantique de la forme :

grondement/NOM==>

[[*gronder* VERBE] *ment* NOM] (*grondement*/NOM, *gronder*/VERBE)

" (Action - résultat de l'action) de *gronder*"

harnachement/NOM==>

[[*harnacher* VERBE] *ment* NOM] (*harnachement*/NOM, *harnacher*/VERBE)

" (Action - résultat de l'action) de *harnacher*"

Cependant, les informations du TLFi seront utiles ici pour deux objectifs :

- 1) typage sémantique en conformité avec le lexique génératif par le biais d'indices lexicaux tels que *action*, *résultat*, *habitude*, *opération*, *fait*, *faculté*, *procédure*, *manière de + verbe à l'infinitif* qui pourront être associés aux catégories d'éventualités du lexique génératif à savoir /activité/, /accomplissement/, /achèvement/
- 2) détection de la structure argumentale des noms prédictifs, ainsi que celle du verbe de base : une fois détecté les indices lexicaux suscités, les procédures d'extraction d'informations sur le TLFi seront spécialisées pour analyser les définitions, afin de repérer les arguments des prédicats nominaux et de leur base verbale. Par exemple, *gronder* est indiqué comme inergatif et *grondement* comme prenant un argument introduit par *de* (*le grondement du tonnerre*). Par contre, *harnacher* est codé comme un verbe transitif et *harnachement* se construit avec deux arguments, l'agent en *par* et le patient en *de* (*l'harnachement des chevaux*).

3.3.5. Typage sémantique des noms en -eur et -oir

Les noms en -eur et ceux en -oir se distinguent selon qu'ils désignent un instrument ou un individu (l'agent du prédicat représenté par le verbe), en ce qui concerne les noms en -eur; un instrument ou un lieu, en ce qui concerne les noms en -oir. De plus, certains noms sont polysémiques (*hachoir* : instrument pour *hacher* / planchette).. Pour l'ambiguïté instrument/individu, les procédures s'appuieront sur une analyse des définitions visant à mettre à jour et à utiliser des indices lexicaux tels que *personne qui*, *ouvrier qui*, *dispositif qui*, *machine qui*, *instrument qui*, *celui qui* etc.

AGRAFEUSE

Machine qui sert à fixer ensemble au moyen d'une agrafe plusieurs feuilles de papiers ou des emballages divers

Ouvrière qui réalise des emballages en carton dont les éléments sont juxtaposés par des agrafes métalliques qu'elle est chargée de fixer

AMORTISSEUR

Dispositif qui atténue la violence de quelque chose

AJUSTEUR

Ouvrier qui ajuste les monnaies

Ouvrier qui met les lames à découper à l'épaisseur voulue

Ouvrier qui trace et façonne des métaux à la main, dans la plupart des cas, et à l'aide d'outils appropriés d'après un plan pour en faire des pièces mécaniques

La même stratégie sera utilisée pour classer les noms en -oir(e).

5. Bibliographie

Anderson, Stephen. 1992. *A-morphous morphology*. Cambridge Studies in Linguistics.

- Cambridge: Cambridge University Press.
- Aronoff, Mark. 1994. *Morphology by Itself*. Cambridge: MIT Press.
- Bassac, Christian, and Bouillon, Pierrette. 2005. Qualia Structure and Anaphoric References in Compounds. Paper presented at *Third International Workshop on Generative Approaches to the Lexicon*, Geneva:27-35.
- Bassac, Christian. 2006. Morphologie et Information Lexicale, Mémoire d'Habilitation à diriger des recherches, Université Michel de Montaigne Bordeaux 3.
- Bouillon, Pierrette. 1997. Polymorphie et sémantique lexicale : le cas des adjectifs., Université Paris 7.
- Busa, Federica. 1997. The Semantics of Agentive Nominals in the Generative Lexicon. In *Predicative Forms in Natural Language*, ed. Patrick Saint-Dizier. Amsterdam: Kluwer.
- Dowty, David R. 1991. Thematic proto-roles and argument selection. *Language* 67:547-619.
- Fradin, Bernard. 2003. *Nouvelles approches en morphologie*. Paris: Presses Universitaires de France.
- Jacquey, Evelyne, and Namer, Fiammetta. 2007. Morphosémantique et modélisation : le cas des verbes dénominaux préfixés par é-. In *Représentation du sens linguistique - Actes du Colloque international de Montréal (2003)*, eds. Denis Bouchard, Ivan Evrard and Etleva Vocaj. Bruxelles: De Boeck / Duculot.
- Johnston, Marc, and Busa, Federica. 1996. Qualia Structure and the Compositional Interpretation of Compounds. Paper presented at *Proceedings of SIGLEX Workshop on Depth and Breadth of Semantic Lexicons (June 22 1996)*, Santa-Cruz, CA.
- Matthews, Peter H. 1991. *Morphology (1st edition : 1974)*. Cambridge: Cambridge University.
- Namer, Fiammetta. 2002. Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. Paper presented at *Actes de Traitement Automatique du Langage Naturel (TALN) 2002*, Nancy, France:235-244.
- Namer, Fiammetta. 2003. Automatiser l'analyse morpho-sémantique non affixale : le système DériF In *Cahiers de Grammaire*, eds. Nabil Hathout, Michel Roché and Nicole Serna, 31-48. Toulouse: ERSS.
- Namer, Fiammetta, and Jacquey, Evelyne. 2003. Lexical Semantics and derivational morphology: the case of the popular é- prefixation in French Paper presented at *GL 2003 : 2nd International Workshop on Generative Approaches to the Lexicon (May, 15-17 2003)*, Geneva:115-122.
- Namer, Fiammetta. 2005. La Morphologie Constructionnelle du Français et les Propriétés Sémantiques du Lexique - Mémoire présenté dans le cadre de l'habilitation à diriger des recherches, UFR Sciences du Langage, Université de Nancy2.
- Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

