INSTITUT DE LINGUISTIQUE FRANÇAISE (ILF)

APPEL D'OFFRES INTERNE 2003 : « SEMANTIQUE »

Exploitation automatique d'informations à partir d'un dictionnaire (le TLFi) pour une représentation homogène et hiérarchisée des informations

Responsables pressentis : E. Jacquey et J-M. Pierrel

Equipe pilote pressentie : Laboratoire ATILF - CNRS
Pascale Bernard, Jacques Dendien, Sébastien Haton, Evelyne Jacquey, Christiane Jadelot, Jean-Yves Kerveillant, Laurence Kister, Josette Lecomte, Nicolas Louis, Jean-Marie Pierrel, Gérard Reb, Philippe Saint-Gérand, Susanne Salmon-Alt, Gilles Souvay

Collaborateurs pressentis[1]
Michael Zock et Patrick Paroubek (LIMSI)
Sylvie Mellet (*Bases, Corpus et Langage (BCL) / UMR 6039 CNRS*, Nice, ILF)
Didier Bourrigault (*Equipe de Recherche en Syntaxe et en Sémantique (ERSS) UMR 5610 CNRS*, Toulouse, ILF)

Organigramme fonctionnel pressenti du projet (cf. page 5)

1. Objectifs et motivations du projet

Que l'on se place dans le domaine de la linguistique théorique (descriptions et modélisations de phénomènes linguistiques) ou dans celui de la linguistique appliquée (didactique, traitement automatique des langues, dictionnairique), les informations lexicales sont très souvent un élément fondamental. Or, une ressource gratuite proposant des informations lexicales de tout niveau n'existe pas encore pour le français. Cette lacune est particulièrement marquée pour ce qui est des informations lexicales sémantiques. Dans la suite, nous appellerons « lexique » une ressource contenant des informations lexicales de tout niveau, et « lexique sémantique » lorsque les informations lexicales seront d'ordre sémantique.

Dans le domaine de la linguistique théorique, l'importance des informations lexicales s'explique par la position centrale du lexique dans les théories. Celui-ci est à la fois le réceptacle de beaucoup d'informations appartenant à différents niveaux de description (phonétique, morphologie, syntaxe, sémantique) et le point d'ancrage de nombre de régularités

[1] Les discussions sur l'implication précise de l'équipe ERSS à Toulouse et de l'équipe « Bases Corpus et Langage » à Nice n'ont pu être entièrement finalisées faute de temps.

modélisées par des règles, des principes, leurs modalités d'applications, etc. (l'ensemble des règles d'une grammaire, les principes de composition sémantique, etc.).

Or, dans ce domaine, les travaux présupposent le plus souvent un lexique pertinent et de couverture suffisante étant donné le phénomène étudié. Ce présupposé n'est cependant pas toujours validé par les faits.

Le coût humain de la constitution d'un lexique satisfaisant, étant donné un phénomène, est important. De plus, les lexiques construits en linguistique théorique s'inscrivent souvent dans l'une ou l'autre théorie linguistique, ce qui ne permet pas une utilisation aisée du lexique ainsi construit dans une autre théorie. Ce serait le cas, par exemple, si un chercheur voulait étudier un phénomène dans le cadre de la théorie du gouvernement et du liage (GB) en s'appuyant sur le lexique développé dans le cadre de la grammaire d'arbres adjoints (TAG). Il en serait de même en sens inverse : comment adapter un lexique élaboré dans le cadre de GB pour s'en servir dans l'étude d'un phénomène dans le cadre de TAG ?

Dans les domaines de la linguistique appliquée, les informations lexicales sont là encore essentielles.

Dans le domaine de la construction d'interfaces de dialogue homme-machine, le lexique constitue le point de jonction entre les mots de la langue, dans laquelle l'interface de dialogue reçoit les requêtes de ses utilisateurs, et l'ensemble des objets et procédures dont dispose le logiciel gouverné par l'interface en question. Or, dans ce domaine-ci, l'une des difficultés récurrentes est la non-réutilisabilité des lexiques construits. Ceux-ci sont souvent construits relativement aux nécessités de l'application visée. Comme il n'existe pas de ressource accessible contenant des informations lexicales de tous niveaux sur le français, les lexiques dans ce domaine du traitement automatique des langues, bien souvent, ne sont pas construits à partir d'études préalables et s'améliorent donc peu au cours des constructions successives.

Il en est de même pour un autre domaine du traitement automatique des langues : la résolution de la référence. En dehors des cas où des calculs statistiques et des régularités permettent de résoudre la référence des pronoms dans les énoncés[2], de même avec la résolution des expressions anaphoriques définies et démonstratives non directes[3], il est nécessaire de recourir au contenu lexical sémantique des mots en présence et aux relations sémantiques entre eux. Mais à nouveau, l'absence d'un lexique sémantique du français conduit les chercheurs à produire des lexiques *ad hoc*.

Enfin, dans le domaine de la dictionnairique (élaboration de dictionnaires), tous les chercheurs s'accordent sur la richesse extraordinaire des informations contenues par les dictionnaires, mais aussi sur l'hétérogénéité de ces informations et sur celle de leur structuration selon les ouvrages. Disposer d'une ressource lexicale sémantique du français fournirait une base de comparaison entre les dictionnaires disponibles et permettrait donc d'intégrer leurs avantages respectifs. De plus, l'absence de ressource lexicale sémantique rend l'élaboration de nouveaux dictionnaires extrêmement coûteuse en temps et en main-d'œuvre.

---

[2] Des anaphores comme *<*Marie*> aime le chocolat et <*elle*> ne s'en prive pas* sont résolues en s'appuyant sur les propriétés morpho-syntaxiques du pronom et de son antécédent. Une phrase comme *<*Marie*> aime la glace et <*elle*> ne s'en prive pas* demande en plus de prendre en compte la similitude de position syntaxique entre le pronom et son antécédent. Enfin, la phrase *Marie aime <*la glace*>, surtout lorsqu'<*elle*> est à la fraise* demande la prise en compte des contenus lexicaux de l'antécédent et de la construction *être à la fraise*.

[3] Une anaphore comme *<*Un homme*> est entré dans la pièce. <*(Cet/L') homme*> portait un chapeau* est ce qu'on appelle une anaphore démonstrative (déterminant démonstratif) ou définie (déterminant défini) directe car il y a reprise exacte du nom régissant, le nom *homme*. De telles anaphores sont actuellement résolues par les algorithmes de résolution de la référence dans les énoncés. En revanche, avec les anaphores non directes, l'appel au contenu sémantique des mots est indispensable : *<*Un homme*> est entré dans la pièce. <*(Cet/L') individu*> portait un chapeau* ou encore *La panne de <*Super phénix*> , <* le surgénérateur de Creys-Malville*> , n' est toujours pas réparée . […] A rajouter aux quelque vingt milliards correspondant au coût de <*la construction*> et aux frais résultant de la fermeture durable du réacteur , après seulement quelques mois de fonctionnement à plein rendement*.

Or, du point de vue de la didactique du français, il serait intéressant de disposer de dictionnaires remplissant des objectifs différents selon le public, le registre et le niveau attendus.

Que ce soit du point de vue de la linguistique théorique ou appliquée, il apparaît donc nécessaire de réfléchir à la constitution d'une sorte de « meta-lexique », notamment sémantique, dont l'objectif principal serait de pouvoir être facilement adapté aux différentes attentes des communautés scientifiques s'intéressant au français (linguistique théorique, lexicographie et dictionnairique, interprétariat, traitement automatique des langues, etc.). Pour ce faire, elle devra idéalement s'appuyer sur toutes sortes de descriptions du contenu sémantique des mots en français.

Les ressources envisageables sont de trois grands types : des lexiques, des modélisations sémantiques de fragments du français et des dictionnaires.

Il n'existe pas de lexique du français accessible. Les lexiques en ligne généralistes (par exemple le Lexique-Grammaire du LADL) ne sont gratuits qu'en consultation, ou alors ils ne contiennent pas toutes les informations que l'on pourrait attendre d'un lexique. C'est le cas dans le lexique MULTEXT, dont une partie est disponible gratuitement sur le serveur ABU, et qui contient des informations flexionnelles et catégorielles mais aucune information liée à la valence des lexèmes ou à leur sens.

Restent donc les deux autres types de ressources : les dictionnaires électroniques et les modélisations sémantiques de fragments du français (parmi d'autres, Patrick Saint-Dizier, François Rastier, Danièle Godard et Jacques Jayez, Georges Kleiber, Pierrette Bouillon, etc.[4]). Les modélisations de fragments de langue, quel que soit le modèle dans lequel ils s'inscrivent, sont difficiles à généraliser à la description du comportement sémantique des mots d'une langue dans son ensemble. Les dictionnaires quant à eux ont d'emblée une visée généraliste. En revanche, la question est souvent posée des modalités d'exploitation des informations qu'ils contiennent.

Parce qu'il nous semble plus accessible de partir d'une description la plus complète possible et de réfléchir à la meilleure manière de l'exploiter, nous prenons le parti de nous appuyer sur un dictionnaire. Parmi les dictionnaires électroniques du français, nous avons choisi le Trésor de la Langue Française Informatisé (TLFi) pour les qualités distinctives qui le caractérisent (taille, couverture, finesse des informations, appui sur un corpus regroupant beaucoup d'œuvres littéraires dans différents genres, possibilités offertes par son informatisation)[«Le Trésor de la Langue Française informatisé: un exemple d'informatisation d'un dictionnaire de langue de référence», Sousmis à TAL, Jacques Dendien, Jean-Marie Pierrel ; Computerized linguistic resources of the research laboratory ATILF for lexical and textual analysis: Frantext, TLFi, and the software Stella, LREC 2002, Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas, Espagne, Vol 3 p. 1090-1098, 27 May - 2 june 2002 Pascale Bernard, Josette Lecomte, Jacques Dendien, Jean-Marie Pierrel ; **Les ressources de l'ATILF pour l'analyse lexicale et textuelle : TLFi, Frantext et le logiciel Stella**, 8èmes Journées Internationales d'Analyse Statistique des Données Textuelles JADT 2002, pages 137-149 Pascale Bernard, Jacques Dendien, Josette Lecomte, Jean-Marie Pierrel].

Ce projet passe donc par une exploitation et une valorisation des données exceptionnelles réunies dans le Trésor de la Langue Française Informatisé (TLFi). Derrière ce premier objectif, la perspective de construire une ressource lexicale du français émerge naturellement. Basée sur le matériau fourni par le TLFi, la ressource lexicale envisagée serait suffisamment riche et générale pour s'adapter aux diverses attentes de différentes communautés scientifiques intéressées. Enfin, la construction d'une telle ressource lexicale sémantique en français s'inscrit dans le projet plus vaste d'élaborer des modélisations en sémantique lexicale qui puissent couvrir une langue dans son ensemble, et dans le cas du projet proposé ici, du français.

Outre la construction d'un lexique sémantique du français, trois participants au moins au projet désirent utiliser les résultats, finaux ou intermédiaires, du projet dans différents domaines de la linguistique théorique ou appliquée. Ce dernier aspect fournira donc un mode d'évaluation au projet, en particulier pour ce qui de l'adaptabilité des résultats obtenus. Michael Zock a pour ambition de réaliser des dictionnaires plus « coopératifs » avec leurs utilisateurs, notamment lorsque ceux-ci sont des apprenants du français, Susanne Salmon-Alt utilisera les résultats de ce projet dans le cadre de la résolution de la référence, en particulier avec les reprises définies et démonstratives non directes et Evelyne Jacquey s'appuiera sur les résultats obtenus pour évaluer les hypothèses attachées à la notion de polysémie logique en français et pourra apporter un élément important dans le cadre de sa collaboration à l'ARC-INRIA GENI (resp. Claire Gardent) dont l'un des objectifs est de s'appuyer sur des informations lexicales riches et structurées pour générer automatiquement des inférences.

Du point de vue du laboratoire pilote pressenti, le laboratoire d'Analyse et du Traitement Informatisé de la Langue Française (ATILF), unité mixte de recherche (UMR 7118) du Centre National de la Recherche Scientifique (CNRS) et membre de l'Institut de Linguistique Française (ILF), un tel projet est en continuité avec la mission passée de l'Institut National de la Langue Française (INaLF), laboratoire du CNRS, mais représente aussi l'évolution culturelle de ce laboratoire. S'appuyant sur l'expérience, le savoir-faire et la culture des membres de l'INaLF, et associant les jeunes chercheurs recrutés en section Sciences et Techniques de l'Information et de la Communication (STIC), ce projet permet de faire collaborer des lexicographes experts, les informaticiens ayant participé à l'informatisation du TLF et des chercheurs pluridisciplinaires, linguistique et informatique, et donc de faire en sorte que la finesse et la richesse des informations du TLF soient valorisées au-delà des seules communautés de la lexicographie et de la linguistique, c'est-à-dire auprès de communautés plus proches du traitement automatique des langues, de la dictionnairique et de la didactique des langues. On peut donc espérer qu'une telle évolution ira dans le sens d'une amélioration non négligeable des approches et des résultats des domaines de recherche relevant de la linguistique appliquée.

Plus globalement, les résultats obtenus dans le cadre de ce projet seront mis à disposition de la communauté de l'ILF, mais aussi des communautés travaillant sur les langues et le langage. Ce dernier aspect aura aussi pour objectif de générer une discussion riche entre les fournisseurs de l'information lexicale (les communautés ayant travaillé ou travaillant encore à l'élaboration de lexiques et celles s'intéressant à la modélisation des informations lexicales sémantiques) et ses destinataires (les communautés de la linguistique théorique, de la lexicographie et de la linguistique appliquée).

2. Descriptif du projet

La ressource lexicale sémantique et la modélisation en sémantique lexicale visées doivent être les plus générales possibles de manière à pouvoir répondre aux attentes de la communauté, attentes qui sont très diversifiées comme nous l'avons évoqué ci-dessus. Pour ce faire, nous avons donc choisi de partir des la richesse extraordinaire des informations contenues dans le TLFi.

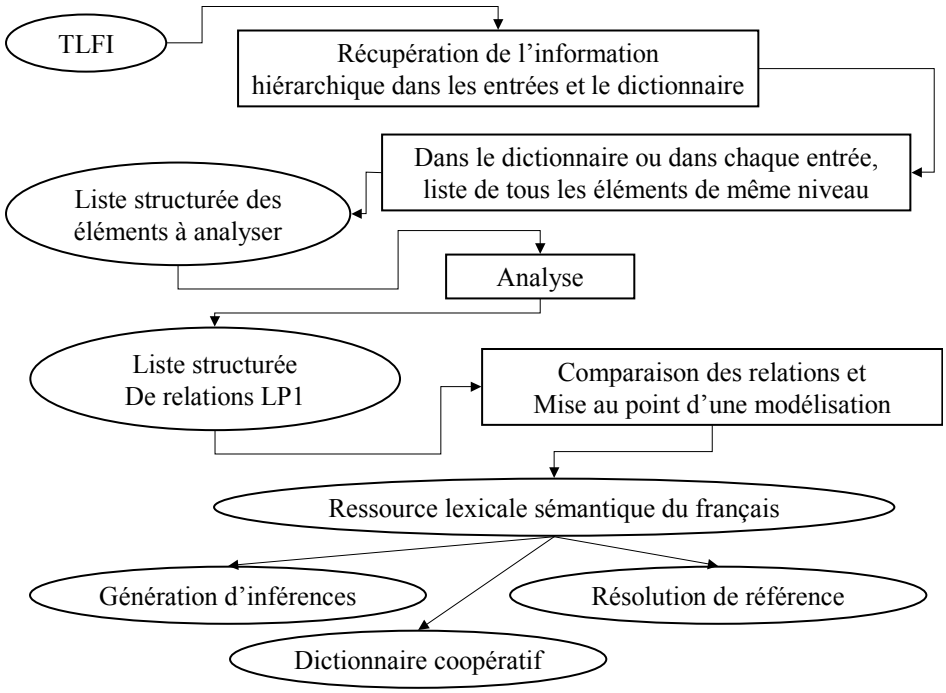
La première question concernera donc en priorité la ou les méthodes, d'extraction et de stockage, des informations fournies par le TLFi. Etant donnée la masse considérable d'informations en jeu, une extraction manuelle et empirique représenterait certainement un coût humain comparable à celui qu'a représenté l'élaboration du TLF lui-même. Il apparaît donc important de mettre au point une méthode d'extraction automatisable à terme. Cette méthode, encore en gestation, constituera l'une des premières phases du projet.

Si l'on observe un dictionnaire de manière très générale, on remarque que les

informations sont soit distinguées, soit regroupées. Les distinctions ou regroupements se situent entre des entrées, des sens, des conditions d'usage ou des indications d'emploi (métonymie, spécialement, technique, figuré, etc). Avec l'aide d'une plate-forme informatique (outils de l'ensemble TLFi – FRANTEXT, étiqueteur WINBRILL, analyseur de G. Reb et N. Louis) pour étudier les distinctions et les regroupements à l'intérieur du TLFi, nous espérons mettre à jour les éléments nécessaires à l'élaboration d'une modélisation du contenu sémantique des mots du français, et par ce biais, à la construction d'une ressource fournissant des informations lexicales sémantiques sur le français.

Dans la suite, nous abordons les principaux aspects du projet. La section (2.1) commente l'organigramme du projet. La section (2.2) décrit rapidement les éléments majeurs du TLFi. La section (2.3) est consacrée à l'analyseur de M. Reb. La section (2.4) est centrée sur un descriptif des applications directes du travail développé dans ce projet pour ses participants.

2.1 Organigramme du projet



Extraction d'un fragment représentatif du TLFi	J. Lecomte, P. Bernard, J. Dendien, C. Jadelot, E. Jacquey
Récupération de l'information hiérarchique dans les entrées et dans le dictionnaire	J. Dendien, J-Y. Kerveillant, J. Lecomte, P. Bernard, E. Jacquey, J-M. Pierrel, ERSS, BCL, M. Zock, P. Paroubek
Analyse des éléments hiérarchiques de même niveau et de même ancêtre	G. Reb, N. Louis, E. Jacquey, C. Jadelot, ERSS, BCL, M. Zock, P. Paroubek
Comparaison des résultats d'analyse sous la forme de relations en logique des prédicats du premier ordre	E. Jacquey, J. Dendien, S. Haton, L. Kister, S. Salmon-Alt, J-M. Pierrel, ERSS, BCL, M. Zock, P. Paroubek
Elaboration d'une modélisation sémantique lexicale pour les éléments analysés	E. Jacquey, J-M. Pierrel, ERSS, BCL, M. Zock, P. Paroubek
Constitution d'une ressource lexicale sémantique du français selon les perspectives de chaque participant	E. Jacquey, J-M. Pierrel, S. Salmon-Alt, ERSS, BCL, M. Zock, P. Paroubek

Le TLFi contient différentes sortes d'informations hiérarchisées. Ces informations peuvent donc être de profondeur différente. Lors de la consultation électronique, l'information hiérarchique est implicite. L'une des tâches de ce projet consiste à la rendre explicite de manière à pouvoir établir des listes d'informations de même niveau. A titre d'exemple, pour déterminer la manière dont les sens d'un même mot sont subdivisés, il est nécessaire d'isoler toutes les définitions de même niveau de ce mot. Divers membres du projet collaboreront à cette tâche, en particulier J. Dendien qui a conçu l'interface de consultation du TLFi, Josette Lecomte et Pascale Bernard qui en connaissent parfaitement toutes les possibilités.

Une fois ces listes établies, que ce soit pour tout le dictionnaire ou à l'intérieur d'une même entrée, leurs éléments doivent être analysés. Les informations de même nature sont rédigées en langue naturelle et leur comparaison sous cette forme est complexe. De plus, les informations appartenant à un même niveau de profondeur au sein d'une entrée ne sont pas obligatoirement de même nature : dans l'entrée de *vue* par exemple, le premier niveau de profondeur contient un crochet --A.- [Gén. avec un art. déf. ou poss.]-- et deux indicateurs d'emploi --B.- par métonymie-- et --C.- Au figuré--). L'analyseur de G. Reb et N. Louis peut transformer les énoncés en listes de relations sémantiques en logique des prédicats du premier ordre (LP1). Après une phase de tests, les résultats sont probants pour les types d'informations du TLFi qui nous intéressent particulièrement, notamment, les conditions d'usage (crochets, constructions, indicateurs d'emploi), les définitions, les syntagmes typiques et les exemples.

A partir des informations de même nature représentées sous la forme de listes de relations sémantiques en LP1, il faudra les comparer pour répondre à plusieurs questions :

- ✓ Quels sont les éléments communs à toutes les informations de même niveau ?
- ✓ Quels sont les éléments qui différencient une information d'autres informations de même niveau ?
- ✓ Etant donnée une liste d'informations de même niveau, qu'ont-elles de commun avec les informations qui peuvent être considérées comme leur ancêtre direct ?

Répondre à ces questions demandera l'étude détaillée des listes de dépendances obtenues en fonction du type des informations de même niveau (définitions, conditions d'usage, indications d'emploi, syntagmes caractéristiques, etc.). La contribution de lexicographes ayant participé à la rédaction des articles, de linguistiques et d'informaticiens s'avérera nécessaire à ce stade. De plus l'analyseur de G. Reb et N. Louis dispose d'un moteur d'inférences permettant de classer les relations sémantiques obtenues après analyse. Nous disposons donc des compétences nécessaires du point de vue théorique et du point de vue pratique.

Dans un deuxième temps, les listes de relations sémantiques ainsi que la liste des propriétés communes et distinctives propres à chaque niveau hiérarchique au sein d'une entrée et au sein du dictionnaire devront être modélisées afin d'obtenir une description homogène du contenu sémantique de tous les mots du dictionnaire.

Enfin, une synthèse des informations modélisées permettra de construire une ressource à même de fournir des informations lexicales sémantiques et capable de s'adapter aux attentes des différentes communautés scientifiques s'intéressant aux langues et au langage.

Nous détaillons à présent les différents éléments ou phases du projet.

2.2 Le Trésor de la Langue Française Informatisé : une mine d'informations

2.2.1 Le Trésor de la Langue Française, version papier

Ce dictionnaire du français a été réalisé entre 1960 et 1994 par l'INaLF, un laboratoire

du CNRS. Edité alors sous la forme de seize volumes, il contient une description lexicographique d'environ 100000 mots avec un total de 270000 définitions. Déjà sous forme papier, de par son ancrage sur des corpus réels (la base FRANTEXT est constituée majoritairement d'œuvres littéraires en français, mais aussi d'autres types d'écrits comme des traités, des essais, etc.[Histoire de Frantex: constitution d'une base textuelle (1964-2002) et perspectives, in *L'édition électronique en littérature et dictionnairique: évaluation et bilan*, Editions Champion, à Paraître, Charles Bernet et Jean-Marie Pierrel]), mais aussi grâce à la synthèse de nombreux documents bibliographiques (notamment d'autres dictionnaires), ce dictionnaire fournit un concentré très riche et très précis des études lexicographiques du français. Outre les définitions, les articles de dictionnaires contiennent des indications d'emploi (syntaxiques, sémantiques, pragmatiques, etc.), des liens de synonymie et d'antonymie, des exemples, principalement extraits de la base FRANTEXT et leur source, et enfin, une étude de l'histoire de chaque mot et de chacun de ses sens (dates et références des premières attestations). Entre 1993 et 2001, l'INaLF a réalisé l'informatisation du TLF. Outre le contenu déjà remarquable de ce dictionnaire, l'informatisation a eu, entre autres, deux avantages majeurs : (1) la communauté scientifique dispose aujourd'hui d'une ressource électronique du français de grande taille et d'une très grande richesse informative et (2) un effort important a été fait pour augmenter l'efficacité de sa consultation.

2.2.2 Les apports de l'informatisation du TLF

La consultation du TLFi (http://www.inalf.fr/tlfi) est effectuée via une interface qui utilise le moteur de recherche STELLA et la théorie des objets complexes [La structuration du TLFi, codage et balisage in *L'édition électronique en littérature et dictionnairique: évaluation et bilan*, Editions Champion, à Paraître, Pascale Bernard, Jacques Dendien, Josette Lecomte, Jean-Marie Pierrel].

❖ *Théorie des objets complexes*

Chaque élément distinguable typographiquement dans un article de dictionnaire a été associé à un objet donné. Le moteur de recherche STELLA a donc analysé les articles de dictionnaire nus et a permis de repérer et baliser les différents objets présents dans chaque article. Le TLFi est ainsi interrogeable sur des types d'objets (**`entrée`**, **`entrée principale`**, **`entrée secondaire`**, **`mot-vedette`**, **`code grammatical`**, **`définition`**, **`texte de définition`**, **`exemple`**, **`indicateur d'emploi`**, **`plan d'article`**, **`construction`**, **`syntagme`**, **`syntagme défini`**, **`syntagme enchaîné`**, **`crochets`**, etc.) et sur le contenu de ces objets, par exemple « rechercher les **`entrées principales`** contenant la séquence *vue* ». On obtient alors tous les articles de dictionnaire du TLFi contenant *vue* dans leur entrée, c'est-à-dire des substantifs composés comme *garde-à-vue*, des locutions comme *à vue*, des adjectifs comme *vu, vue* et aussi, bien sûr, le substantif *vue*.

Pour permettre des recherches plus complexes, les objets sont reliés entre eux par des liens d'inclusion ou des liens de dépendance. Ceci permet d'effectuer des requêtes telles que « rechercher les **`entrées principales`** contenant la séquence *vue* et, inclus dans ce premier type d'objet, les **`codes grammaticaux`** contenant la séquence *subst* ». Ainsi, les résultats sont restreints au seul article du substantif *vue*. Une autre requête intéressante peut être de chercher toutes les définitions du substantif *vue*, ce qui se traduirait par « rechercher les **`entrées principales`** contenant la séquence *vue*, incluant un **`code grammatical`** contenant la séquence *subst*, et dépendant des **`entrées principales`**, les objets **`définitions`** dans lesquels seraient inclus des objets **`textes de définitions`** ».

On se représente aisément le potentiel qu'offre la théorie des objets complexes pour affiner les requêtes. De plus, chaque type d'objets contient des informations de nature diverse. Parmi la diversité des objets du TLFi, certains ont particulièrement retenu notre attention dans la perspective d'une exploitation des informations présentes dans le dictionnaire. L'objet

construction indique les conditions syntaxiques d'emploi d'un mot et ce type d'objet est souvent accompagné d'une **définition**, de syntagmes caractéristiques (**syntagme, syntagme défini, syntagme enchaîné**) et d'**exemples**. Il apparaît donc évident que le TLFi peut constituer un matériau fondamental dans la perspective d'élaborer une ressource lexicale du français, adaptable aux attentes de communautés scientifiques distinctes.

❖ *Deux modes d'affichage : global ou détaillé*

Tous les résultats de requêtes peuvent être affichés de manière globale ou détaillée.

L'affichage global montre tout d'abord le nombre de résultats et les articles dans lesquels ces résultats ont été trouvés. Ensuite, ce type d'affichage peut être paramétré selon les types d'objets que l'on veut voir apparaître à l'écran. Ce premier type d'affichage permet donc, comme son nom l'indique, de se faire une idée synthétique des résultats d'une requête.

L'affichage détaillé consiste à afficher l'article complet dans lequel a été trouvé un résultat. Ce deuxième type d'affichage permet donc de vérifier la correction et la précision des résultats, notamment en termes de degré de dépendance entre les objets demandés lors de la requête.

2.2.3 Extraction d'informations et modalités de requêtes dans le TLFi : questions non résolues

Bien que l'interface du TLFi et la théorie des objets complexes apportent une amélioration notoire du degré de complétude et de précision dans la consultation du TLFi, certaines difficultés émergent.

❖ *Objets du TLFi et diversité de leur contenu*

Nous prendrons à titre d'illustration deux exemples : les **indicateurs d'emploi** et les **crochets**.

De nombreux chercheurs ont élaborés des classifications de la polysémie, en français et dans d'autres langues. Il serait intéressant de pouvoir les comparer en s'appuyant sur les articles du TLF. Un premier examen montre, concernant les classes de polysèmes logiques de (Pustejovsky 1995), que des sens appartenant à une même classe de polysémie logique, n'apparaissent pas forcément selon une structure comparable dans les articles de dictionnaire. Les noms *bouteille* et *livre*, par exemple, sont considérés par Pustejovsky comme appartenant à la classe « contenant/contenu ». Dans les articles correspondants, le sens de « contenu » est considéré comme une métonymie avec le nom *bouteille*, ce qui donne lieu à l'utilisation d'un **indicateur d'emploi**, alors qu'il donne lieu à une subdivision non marquée par un quelconque indicateur d'emploi avec le nom *livre*. Par ailleurs, même si l'indication d'une métonymie semble majoritaire, elle n'est pas systématique. Enfin, l'indication de métonymie concerne aussi d'autres sens, plus ou moins spécifiques, plus ou moins figés, qui n'appartiennent pas aux théories lexicales ayant étudié la polysémie logique. Cela n'est pas étonnant car la lexicographie vise une description à la fois complète et aussi précise que possible du sens des mots. En comparaison, les théories lexicales visent elles aussi une description aussi complète et précise que possible, mais en ayant en même temps pour objectif la découverte de régularités et de classes. Les théories lexicales s'intéressent donc souvent, non pas à la langue dans son ensemble et dans toute sa complexité, mais à un fragment de celle-ci pour lequel elles tentent d'établir des classifications, des régularités permettant de produire des modélisations. Par conséquent, une étude fine du contenu des **indicateurs d'emplois** s'avère nécessaire car elle enrichirait les théories lexicales tout en apportant une homogénéisation et une systématisation de la structure lexicographique, ce qui pourrait par exemple s'appliquer aux sens des mots polysémiques.

Un autre exemple est celui des **crochets**. Une simple requête sur ceux qui dépendent du mot-vedette dans l'article consacré au substantif *vue* montre une grande diversité de contenu (conditions syntaxiques, sémantiques et/ou pragmatiques) et de forme.

VUE, subst. Fém
[Gén. avec un art. déf. ou poss.]
[Dans des empl. faisant réf. aux performances de ce sens, à la qualité des perceptions]
[En rapp. avec les capacités moy. ou individuelles de l'œil hum.]
[Avec un adj. ou compl. de nom]
[Dans des empl. faisant réf. à l'exercice de ce sens]
[Relativement à l'objet vu]
[Avec prop. indiquant la conséquence]

A nouveau, cet état de fait s'explique par la définition du contenu d'un objet de type **crochet** : un tel objet contient des informations sur les conditions d'usage du mot-vedette, conditions qui peuvent donc être de natures variées. L'analyse du contenu de ces objets viendra enrichir la finesse des informations extraites à partir du TLFi, analyse qui est rendue possible par l'analyse de G. Reb et N. Louis et exploitable par le biais du moteur d'inférences associé.

❖ *Degré de dépendance et correction des résultats d'une requête*

La théorie des objets complexes associée au formulaire des recherches complexes dans l'interface du TLFi permet de faire des requêtes sur six types d'objets simultanément, chacun d'eux pouvant être lié par des liens d'inclusion ou de dépendance aux autres types d'objets. Cette fonctionnalité permet donc de construire des requêtes très complexes.

La difficulté qui apparaît cependant concerne la correction des résultats obtenus. Un exemple représentatif est une requête cherchant à associer aux **crochets** de l'exemple du substantif *vue*, les **définitions** et **liens de synonymie/antonymie** pertinents. Aucun lien d'inclusion n'associe ces objets dans la théorie des objets complexes. Restent les liens de dépendance. Cependant, ces liens n'étant pas directs, deux stratégies extrêmes sont envisageables. Si l'on applique une dépendance lâche, tous les objets doivent uniquement être dépendants de l'**entrée principale**, les résultats sont très bruités (de l'ordre de 13000 résultats alors que l'article de *vue* ne contient que huit instances de l'objet **crochets**). Si l'on applique une dépendance stricte, les objets dépendent les uns des autres deux à deux (l'objet **crochets** dépend de l'objet **entrée principale**, l'objet **définitions** dépend de l'objet **crochets** et l'objet **lien de synonymie/antonymie** dépend de l'objet **définitions**), les résultats ne sont pas complets (seulement quatre résultats dont deux concernent le même objet **crochets**). Une solution possible serait d'examiner tous les liens possibles d'inclusion et de dépendance entre chacun des objets, l'identité qui est une inclusion particulière et le cas où les objets ne sont liés par aucune relation, ce qui conduirait dans cet exemple à 18 requêtes distinctes, soit dans le cas général à N(N-1)*3 requêtes pour un nombre N d'objets. De plus, l'ensemble important de résultats ainsi obtenus doit être filtré pour extraire les dépendances directes. L'exemple suivant compare le premier résultat de la requête avec dépendance stricte ou lâche avec la portion de l'entrée de dictionnaire correspondante. Pour prendre un exemple de discordance, la définition (balises ▪3) ne décrit pas la signification de l'ensemble des emplois correspondant au **crochet**, mais seulement à une partie de ceux-ci, partie qui est associée au sens **A.1**. De plus, cette définition est rapprochée d'un **synonyme** (balise ▪4) alors que celui-ci est en fait lié à la locution *Voir de sa propre vue*, elle-même incluse dans la description d'un second sens du nom *vue*, le sens **A.2**.

VUE, subst Fém.
Résultat de la requête
(**crochets**) ▪2[Gén. avec un art. déf. ou poss.]▪2
(**définitions**) ▪3Sens par lequel on perçoit la lumière, la forme, la couleur, la position des objets; p. méton., fonction remplie par ce sens, perception qui lui est

propre.[3]
(**`lien de synonymie/antonymie`**) [4]Synon. plus cour. voir de ses propres yeux (v. œil II A).[4] **(Voir de sa propre vue.)**

Portion de l'article de dictionnaire

A. [Gén. avec un art. déf. ou poss.]
1. Sens par lequel on perçoit la lumière, la forme, la couleur, la position des objets; p. méton., fonction remplie par ce sens, perception qui lui est propre.
2. [Dans des empl. faisant réf. aux performances de ce sens, à la qualité des perceptions]
a) [En rapp. avec les capacités moy. ou individuelles de l'œil hum.]
-- [Avec un adj. ou compl. de nom]
Voir de sa propre vue. Synon. plus cour. voir de ses propres yeux (v. œil II A).
Etc.

De tels résultats proviennent du fait que les liens de hiérarchie ne sont pas pris en compte de manière explicite dans l'interrogation. Une solution serait d'utiliser des liens d'inclusion pour représenter les liens hiérarchiques. Cependant, une telle approche entrerait en contradiction avec la définition du contenu des objets et de plus, elle rendrait cette définition extrêmement variable concernant l'ordre hiérarchique et la nature des objets inclus. C'est pourquoi l'inclusion n'est utilisée que dans des cas très restreints comme, par exemple, dans le cas de l'objet **`entrée`** dans lequel sont inclus les objets **`mot-vedette`** et **`code grammatical`**. Pour prendre en compte des liens de hiérarchie, une interrogation utilisant l'objet **`plan de l'article`** paraît la plus évidente. Mais à nouveau, le nombre de résultats est décuplé, insuffisant ou requiert l'exécution d'autant de requêtes qu'il y a de combinaisons possibles puis le tri des résultats.

2.2.4 Explicitation des liens hiérarchiques entre objets,

❖ *Analyse fine des lacunes dans les fonctionnalités offertes pour la consultation du TLFi*

Les difficultés rencontrées dans la consultation du TLFi se concentrent sur la précision des résultats obtenus. Comme nous l'avons précisé ci-dessus, ces difficultés reposent sur deux caractéristiques du TLFi : (1) la diversité en contenu et en forme de certains types d'objets et (2) la non-prise en compte explicite des liens de hiérarchie à l'intérieur d'un article de dictionnaire.

- ✓ **Liens de hiérarchie** : Le TLFi existe aussi sous la forme d'un texte balisé. Un examen rigoureux des différentes configurations hiérarchiques possibles entre les objets permettra d'établir une liste de cas.
- ✓ **Diversité en contenu et en forme** : Une fois le problème de la prise en compte des liens de hiérarchie résolu, une étude fine de certains types d'objets, alliant savoir des lexicographes et outils d'analyse automatique, sera menée.

❖ *Mise en œuvre des modifications nécessaires : trois stratégies pour la prise en compte de liens hiérarchiques*

Etant donné que le TLFi existe aussi sous la forme d'une base textuelle balisée, il peut être considéré, non plus comme un dictionnaire, mais comme un corpus d'un type particulier et il est possible de lui appliquer des outils d'interrogation de bases textuelles adaptés, approche qui a d'ailleurs été adoptée par J. Dendien pour la mise au point du moteur d'analyse STELLA et celle de l'interface d'interrogation du TLFi. Une fois la liste de combinaisons hiérarchiques possibles établie dans la phase précédente, trois stratégies seront examinées pour prendre en compte les modifications nécessaires.

- ✓ **Modification des formulaires** : Dans l'interface concernant les recherches complexes, plusieurs modifications seraient nécessaires pour étendre le langage de

requêtes : (1) ajouter la possibilité de représenter le contenu d'un objet comme une variable, ceci permettant d'imposer des partages de valeurs et donc d'assurer de l'identité de deux objets (par exemple, pour s'assurer que deux définitions dépendent bien d'un même niveau de hiérarchie), (2), intégrer la possibilité de paramétrer le degré de dépendance entre deux objets et (3), permettre l'interrogation des balises de hiérarchie, visibles dans la base textuelle mais non interrogeables avec l'interface.

- ✓ **Utilisation des formulaires suivie d'une interrogation de la base textuelle du TLFi** : L'interface du TLFi serait modifiée uniquement pour permettre un rapatriement des résultats sous la forme de textes contenant l'ensemble des balises présentes dans la base textuelle du TLFi. Ensuite, ces fragments d'articles de dictionnaire seraient soumis à des outils d'interrogation de corpus. A ce stade, il sera nécessaire de mettre en place différentes plate-formes d'outils informatiques et de comparer leurs résultats respectifs en termes de correction des résultats (étiqueteur morpho-catégoriel WINBRILL, Gilles Souvay et Josette Lecomte, analyse syntaxique et classification, G. Reb et N. Louis, concordancier MONOCONC, P. Bernard et J-Y. Kerveillant, associé à un outil d'interrogation de corpus analysés syntaxiquement TIGER, J-Y. Kerveillant et E. Jacquey).
- ✓ **Interrogation de la base textuelle du TLFi** : Mise au point et application d'une plate-forme informatique pour l'interrogation d'un tel corpus. Celle-ci reproduirait en partie la philosophie de l'interface du TLFi (degré variable dans la complexité des requêtes) mais en intégrant dès le départ la prise en compte des liens hiérarchiques.

A ce stade, pour chaque niveau à l'intérieur d'une entrée ou au sein du dictionnaire dans son ensemble, nous disposerons, pour chaque niveau de hiérarchie, de :

- ✓ La liste des informations de même niveau (ses frères)
- ✓ La liste de ses ancêtres niveaux supérieurs (sa généalogie)

2.3 Analyse et classification

2.3.1 Présentation succincte de l'analyseur

L'analyseur construit par G. Reb et N. Louis vise l'extraction des contenus propositionnels des phrases analysées. Il repose sur une grammaire caractérisée par plusieurs propriétés.

- ✓ La grammaire est formelle : les théorèmes sont dérivés d'un ensemble de termes et d'axiomes de base par application de règles d'inférence explicites
- ✓ La grammaire est non contextuelle : elle est fondée sur l'association de règles syntaxiques et de règles sémantiques
- ✓ La grammaire est définie par un quadruplet : le vocabulaire terminal (dictionnaire de formes de mots = 54173), le vocabulaire auxiliaire (66 variables grammaticales), les règles de production (une centaine de règles) et un axiome (la phrase)
- ✓ La grammaire est générative de type syntagmatique : elle définit un langage (un ensemble de phrases-systèmes) contrôlé par extension de règles caractérisant une procédure de décision et une description structurelle spécifique (les fonctions syntactico-sémantiques)

L'analyseur a été appliqué à différentes sortes de textes (corpus de presse, textes littéraires, etc.). Le résultat de l'analyse d'une phrase se présente sous deux formes : (1) une analyse syntaxique dans le formalisme de l'analyseur et une liste de relations sémantiques en LP1 dans lesquelles les prédicats logiques sont conformes aux formes prédicatives en langue

du point de vue de la valence de ces dernières.

Le 15 octobre 1969, Hubert Curien succède à Pierre Jacquinot à la direction générale du CNRS.
Relations :
1. [8]succéder(hubert, curien ; pierre, jacquinot)

Actuellement, il a été adapté pour analyser le contenu des objets du TLFi, objets qu'il classe en cinq niveaux :

- ✓ Niveaux 1 et 2 les (mots-vedette + codes grammaticaux) et les définitions
- ✓ Niveau 3 les crochets et les indicateurs d'emploi
- ✓ Niveau 4 les constructions
- ✓ Niveau 5 les exemples et les syntagmes

L'analyseur, appliqué aux objets du TLFi, est à même de produire une représentation unifiée de leur contenu. L'analyseur peut donc extraire le contenu propositionnel de tout objet analysé, sous la forme d'un contenu propositionnel représenté par une ou plusieurs relations en LP1 et adapté au type d'objet analysé (niveaux 1, 2, … ou 5). L'enjeu est ensuite de les exploiter.

2.3.2 Exploitation des contenus propositionnels extraits

Selon G. Reb, les représentations des objets analysées pourraient être exploitées par inférence, conduisant ainsi à deux types de résultats.

- ✓ **Déductions par généralisation** : le traitement unifié des crochets et des indicateurs d'emploi (niveau 3) avec celui des exemples (niveau 5) permet une généralisation par appariement du contenu des crochets ou indicateurs d'emploi avec le syntagme auquel il est fait référence dans l'objet de niveau 3. Les objets de niveau 3 fonctionneraient comme des hyperonymes par rapport aux syntagmes auxquels ils font référence. Or ces syntagmes peuvent s'identifier avec les exemples qui fonctionneraient par transitivité comme des hyponymes par rapport aux objets de niveau 3.

  ETUDIER, verbe
  II.A.1.a. [l'objet désigne une discipline d'enseignement]
  ex : « *Je croyais, dit timidement M. Delteil, qu'il n'y avait pas moyen d'étudier la botanique sans connaître un peu de latin* »
  ATTAQUER, verbe
  I.B.2.b. [le sujet désigne un phénomène naturel ou un agent inanimé]
  ex : « *La rouille attaque le fer, la maladie attaque l'organe* »

  Avec *étudier*, le crochet précise les conditions d'emploi de l'objet du verbe. Dans l'exemple, l'analyse permet d'identifier l'objet de l'instance du verbe *étudier* et le groupe nominal *la botanique*. On peut donc obtenir automatiquement la déduction *la botanique est une discipline d'enseignement*.
  Avec *attaquer*, le principe est le même à ceci près que le constituant servant de variable est cette fois le sujet. Les déductions obtenues sont de la forme *la rouille est un phénomène naturel ou un agent inanimé*, *la maladie est un phénomène naturel ou un objet inanimé*.

- ✓ **Définition systématique des constructions (niveau 4) via l'instanciation des positions syntaxiques par des vocables appartenant aux classes lexicales appropriées** : Dans les constructions, les arguments des formes propositionnelles correspondant à l'analyse sont représentées par les formes postiches *qqn/qqc*. En

utilisant la relation d'hyperonymie/hyponymie, ces formes peuvent être instanciées à partir des définitions, des syntagmes ou des exemples et venir compléter les constructions.

ACCOMMODER, verbe

I.A.1.a. Accommoder qqn de qqc.

Déf : Céder quelque chose à quelqu'un pour de l'argent, vendre

ex : *... nous nous rendons chez un négociant russe, qui nous* ***accommode de*** *deux magnifiques fourrures tout récemment arrivées de Vitinsky. Un Grec, de Smyrne, nous vend quatre schalls de Cachemire, ...* (V. DE JOUY, *L'Hermite de la Chaussée d'Antin,* t. 2, 1812, p. 115)

I.B.2.b. Accommoder qqc. à.

Déf : L'adapter à, la mettre en correspondance avec quelque chose ou plus rarement avec quelqu'un :

ex : *C'est faire tort au catholicisme que de l'****accommoder*** *ainsi* ***à*** *nos idées modernes, outre qu'on ne le fait que par des concessions verbales qui dénotent mauvaise foi ou frivolité. Tout ou rien, les néo-catholiques sont les plus sots de tous.* (E. RENAN, *Souvenirs d'enfance et de jeunesse,* 1883, p. 402.)

A partir de l'association des constructions, des définitions et des exemples, on obtiendrait plusieurs déductions dont, par exemple, *accommoder quelqu'un de quelque chose est équivalent à céder quelque chose à quelqu'un ou vendre quelque chose à quelqu'un*, ou encore *accommoder quelque chose à quelque chose est équivalent à adapter quelque chose à quelque chose ou plus rarement quelqu'un ou mettre en correspondance quelque chose avec quelque chose ou plus rarement quelqu'un*.

L'intérêt d'un tel système d'inférences est de construire une base de données dans laquelle seront regroupés tous les contenus propositionnels équivalents.

2.3.3 Mise en œuvre dans le projet

Avant de procéder à l'analyse automatique des informations de même niveau hiérarchique, nous établirons une liste de configurations caractéristiques en fonction de deux paramètres : nature et profondeur des informations appartenant à une même entrée.

Configuration 1 Mot-vedette, code gram. Sub1{déf., synt., ex., … Sub2{crochets, déf., synt., ex., … Sub3{ind. d'emploi, déf., … }Fin_Sub3 }Fin_Sub2 }Fin_Sub1	…	Configuration N Mot-vedette, code grammatical Sub1{crochets, synt., ex., … Sub2{ind. d'emploi, déf., synt., ex., … Sub3{ déf., … }Fin_Sub3 }Fin_Sub2 }Fin_Sub1

A la suite du choix d'une instance de chaque configuration, l'analyse sera lancée pour chaque liste d'informations de même niveau hiérarchique.

Comme premier résultat, nous espérons obtenir les classes d'équivalences entre contenus propositionnels, classes propres à chaque niveau de profondeur dans toutes les instances de configurations prises en compte.

2.4 Exploitation des résultats et applications concrètes

❖ *Modélisation, évaluation d'hypothèses en sémantique lexicale de la polysémie et lexique pour la génération d'inférences*

L'ensemble des classes d'équivalences pour chaque niveau de profondeur permet une approche synthétique du contenu d'une entrée de dictionnaire. Trois applications sont envisageables.

- ✓ **Modélisation d'informations sémantiques applicable au français** : L'intérêt d'une modélisation est très lié à la construction à terme d'une ressource lexicale du français fournissant, outre les informations que l'on peut trouver par exemple dans MULTEXT, des informations sémantiques. Pour être utile, cette ressource sémantique doit être modulaire, permettant ainsi le choix de certains types d'informations et en excluant d'autres ; elle doit être normalisée ; elle doit être compatible ou pouvoir intégrer des ressources traitant d'autres niveaux de description (en particulier, phonétique, morphologique, syntaxique, fréquence d'usage, etc.). L'enjeu à ce stade du projet sera donc de délimiter un mode de représentation le plus expressif possible tout en préservant la possibilité d'une normalisation et d'une traduction aussi simple que possible vers d'autres formalismes demandés par des acteurs de la communauté (Logiques de description, Structures de traits typées, Logique des prédicats du premier ordre, Lambda-calcul étendu, Logique des enregistrements, etc.).
- ✓ **Evaluation d'hypothèses en sémantique lexicale** : Extraction d'un lexique constitué des mots supposés appartenir à la classe de la polysémie logique selon (Pustejovsky 95), à la classe des mots à sens coopérants selon (Cruse 85-86), à la classe des polysèmes avec coprédication selon (Kleiber 99, Godard et Jayez 93) et à la classe de la polysémie constructive selon (Copestake et Briscoe 95). Le premier point va consister à vérifier les hypothèses concernant les sens véhiculés par les mots étudiés dans les travaux de ces chercheurs : les classes d'équivalence font-elles apparaître ces sens, à quels niveaux de profondeur ? Le second point s'intéressera à l'extension de la classe des mots polysémiques avec coprédication en français. Enfin, une modélisation de cette classe pourra être proposée tout en étant ancrée sur une description généraliste du français, à savoir le TLF.
- ✓ **Construction de lexiques dédiés pour le traitement automatique des langues** : Extraction des informations pertinentes pour la construction d'un lexique destinée à la génération automatique d'inférences, collaboration à l'ARC-INRIA GENI (responsable Claire Gardent, équipe pilote Langue et Dialogue, LORIA UMR INRIA – CNRS – Nancy1 – Nancy2). Dans le projet GENI, la génération d'inférences sera prise en charge par le moteur d'inférences RACER. Ce moteur utilise des formules des logiques de description. L'une des questions intéressantes liées à la collaboration avec GENI est la faisabilité de la modélisation d'informations lexicales sémantiques dans le formalisme des logiques de description.

❖ *Exploitation du TLFi pour assister un rédacteur à trouver le mot « qu'il a sur le bout de langue » (M. Zock et P. Paroubek)*

Le but à long-terme de M. Zock et de P. Paroubek (ZP) est de construire une *mémoire associative* afin d'assister le locuteur/rédacteur à trouver le mot que celui-ci a sur le bout de la langue.

Il est évident qu'un dictionnaire est un composant essentiel pour tout système de traitement de la langue (naturel ou artificiel) qu'il s'agisse d'analyse ou de production. Cependant, malgré le nombre de dictionnaires électroniques d'excellente qualité, on est très loin d'en faire un bon usage, c'est-à-dire de les exploiter au mieux. Les possibilités du support électronique sont à la fois sous-exploitées et mal utilisées. Elles sont sous-exploitées dans la mesure où l'on n'utilise pas à fond la puissance et la flexibilité des machines pour traiter l'information contenue dans la base lexicale. Elles sont mal exploitées notamment en ce qui

concerne l'accès et d'affichage. A information égale, il est clair qu'on peut faire beaucoup plus de choses avec un dictionnaire électronique qu'avec un dictionnaire papier.

L'objectif de cette collaboration entre des linguistes, informaticiens et psycholinguistes consiste à concevoir un ensemble de fonctionnalités pour assister l'homme à *traiter* la langue, en exploitant la puissance et la souplesse de l'informatique afin de naviguer dans un vaste réseau d'informations.

Notre objectif est de doter des dictionnaires électroniques d'un certain nombre de fonctionnalités afin d'en augmenter les performances et l'utilité. La démarche se veut générique, elle n'est donc pas limitée aux dictionnaires en question, en l'occurrence le TLFI.

Ce dernier nous sert seulement de test de validation.

Partant d'une certaine conception du dictionnaire mental (mémoire associative où tous les concepts /mots sont hautement interconnectés), nous allons en construire un simulacre afin d'obtenir des performances analogues à celles du dictionnaire mental : l'ordinateur doit deviner le mot ou assister le locuteur à trouver le mot que celui-ci a sur le bout de la langue. Nous allons donc construire un simulacre du dictionnaire mentale pour le tester ensuite sur celui qui nous a servi de modèle, l'homme. La trace des enregistrements d'interactions doit nous révéler les stratégies (de navigation) des utilisateurs en train de chercher des mots. Ainsi nous aurons fermé la boucle entre le *modèle*, son *implémentation* et sa *validation* avec retour possible sur le modèle.

(ZP) ont donc pour but de construire un dictionnaire à la fois riche et facilement consultable pour l'*analyse* et la *production* du langage[5]. A cette fin, ils proposent de construire un dictionnaire analogue à celui des êtres humains, à savoir un dictionnaire, qui, outre les informations conventionnelles (définition, forme écrite, informations grammaticales) contiendrait des liens (associations), permettant de naviguer entre les idées (concepts) et leurs expressions (mots). Un tel dictionnaire permettrait donc l'accès à l'information recherchée soit par la forme (lexicale : analyse)[6], soit par le sens (concepts : production), soit par les deux, simultanément, ou l'un après l'autre[7].

L'intuition selon laquelle le dictionnaire mental (d'aucuns préfèreraient parler d'encyclopédie) serait un réseau, dont les *noeuds* sont des mots (et/ou des concepts), et les *liens* essentiellement des associations ne date pas d'hier. Pourtant, bien que cette intuition ne soit pas nouvelle et bien qu'elle soit partagée par de nombreux chercheurs[8], il n'y a à leur connaissance aucun inventaire (ou de classification) quant à la nature de ces liens[9]. Or, il est clair que, pour pouvoir créer un tel dictionnaire, il faut avoir fait l'inventaire des associations (une des contributions de M. Zock) et organiser les données en conséquence. A cette fin (ZP) poursuivent deux pistes, l'une analytique (chercher dans des ontologies et des thesaurus), l'autre empirique. Dans ce dernier cas ils prévoient deux possibilités :

(a) utiliser des moyens informatiques pour faire des listes de collocations (mots apparaissant ensemble dans une phrase); Dans ce but ils envisagent d'utiliser un outil créé au Limsi à cet effet. Cependant, le résultat sera uniquement une liste de cooccurences, la nature du lien reste

[5] Ce dernier aspect est souvent négligé. Des thesauri comme celui de Roget (1852), de Péchoin (1992) et le *Language Activator* (1993) étant des exceptions notables à cet égard.

[6] Sur cet aspect, on se reportera à (Zock et Fournier, 2002) pour une stratégie d'accès lexical par la forme, cette stratégie ayant été mise en œuvre sous la forme d'un programme informatique.

[7] Il arrive qu'on ne trouve pas d'emblée le bon candidat, auquel on part d'un mot raisonnablement proche (synonyme, hyperonyme, antonyme), espérant par ce biais de se rapprocher du mot cible.

[8] En effet, cette intuition se trouve déjà chez Aristote (« De memoria et reminiscentia »), puis chez des *philosophes* (Locke, Hume) et *physiologistes* anglais (James et Stuart Mills), des *psychologues* (Galton, 1880 ; Freud, 1901 ; Jung & Ricklin, 1906) et des *psycholinguistes* (Deese, 1965 ; Jenkins, 1970). Enfin, cette idée est sousjacente à WORDNET (Miller,1990), aux travaux connexionistes (Stemberger, 1985 ; Dell, 1986), aux hypertextes et au web (Bush, 1945 ; Nelson, 1967). Pour des synthèses en psycholinguistique voir (Hörmann, 1972 ; chapitres 6-10), pour des références plus récentes voir (Spitzer, 1999). Enfin, pour des travaux portant sur des *réseaux sémantiques*, voir (Sowa, 1992).

[9] A cet égard, le travail de Mel'cuk (1992) est peut-etre encore une des meilleurs pistes.

donc à explicitier, d'où l'intérêt de l' approche psycholinguistique (voir ci-dessous)

(b) l'approche psycholinguistique : on demande à des êtres humains de nommer le lien entre des couples de mots. Une variante serait de donner aux gens des mots stimulus auxquels ils doivent répondre par le premier mot leur venant à l'esprit, puis on leur demande d'expliciter la nature du lien, l'hypothèse implicite étant que les gens savent mieux trouver la nature des liens pour les couples de mots produits par eux-mêmes que pour ceux donnés par quelqu'un d'autre.

Enfin, on peut également observer des gens en train de consulter un dictionnaire et leur demander de révéler leur méthode. Que savent-ils à propos d'un mot (partie du sens, relation avec d'autres mots/concepts, origine, nature et nombre de syllabes, etc.)? Quelles sont les questions qu'ils se posent? En d'autres termes, comment font-ils pour trouver l'objet recherché, à savoir, un mot ou une séquence de mots particuliers.

3. Productions envisagées

L'intérêt d'un tel projet est d'avancer sur les manières d'exploiter une ressource telle que le TLF ayant des caractéristiques de taille, de couverture et de richesse qui permettent de la considérer comme une ressource généraliste du français. Dans cette optique, les membres du projet auront à cœur d'élaborer un manuel de référence présentant les méthodes d'extraction en détail à partir des informations fournies par le TLFi et son interface. Ce manuel sera présenté et discuté lors de la seconde réunion plénière du projet.

Par ailleurs, dans l'optique d'un partage des résultats avec les communautés s'intéressant au langage et aux langues, en particulier au français, une interface sera mise en place à l'ATILF pour permettre la consultation libre des résultats du projet en fonction des différentes perspectives de celui-ci.

Enfin, les résultats du projet seront accessibles pour les partenaires via des modes d'exploitation futurs qui seront régis par une convention définie par les partenaires eux-mêmes en fonction de leur investissement respectif.

4. Collaborations inter-équipes envisagées (au sein de l'ILF et hors ILF)

L'équipe de Toulouse, l'ERSS membre de l'ILF, jouit d'une grande expérience dans la manipulation de corpus, que ce soit sur le plan de leur construction, de leur traitement ou de leur interrogation. Par ailleurs, il s'agit d'une équipe expérimentée dans l'art et la manière d'extraire des informations à partir de corpus. Enfin, plusieurs thésards de cette équipe pourront nettement enrichir leurs recherches en collaborant directement à ce projet, notamment pour ce qui est de l'utilisation des données fournies par l'ATILF.

L'équipe du LIMSI, Michael Zock et Patrick Paroubek, sont intéressés à l'élaboration de dictionnaires plus coopératifs. Pour ce faire, ils aimeraient construire un réseau sémantique très riche avec des liens nommés, mais plus variés qu'habituellement. Leur collaboration dans ce projet, outre le partage de leurs nombreuses expériences dans le domaine de la dictionnairique et du traitement automatique des langues, leur permettra de développer l'expérience de la constitution du réseau sémantique qu'ils envisagent en s'appuyant sur l'exploitation des informations présentes dans le TLFi.

L'équipe de Sylvie Mellet, de Nice, membre de l'ILF ?

5. Dimension internationale envisagée

L'ensemble des membres du projet s'engagent bien entendu à divulguer le résultat des recherches menées dans le cadre de conférences et revues, nationales et internationales.

En outre, nous étudions actuellement la possibilité d'une insertion dans le cadre du réseau Lexique, Terminologie et Traduction (LTT) de l'AUF afin de pouvoir impliquer d'autres équipes francophones ou s'intéressant à l'étude du lexique du français et afin d'assurer une diffusion plus large des résultats du projet dans le monde de la francophonie.

6. Calendrier prévisionnel du projet (avec étapes)

Le calendrier est envisagé sur deux ans.

Le départ du projet consistera en une réunion plénière où l'ensemble des participants se répartiront volontairement dans les différentes étapes et les diverses tâches par étape.

La première étape se concentrera sur **l'extraction de listes d'objets** de même niveau étant donné un ensemble d'entrées représentatives dans le dictionnaire. La définition de cet ensemble d'entrées représentatives se fera en étroite collaboration avec les lexicographes du laboratoire pilote mais aussi en concertation avec tous les membres du projet en fonction de leurs attentes précises. A la suite de la délimitation de cet ensemble, trois équipes seront constituées et seront chargées d'expérimenter chacune l'une des trois stratégies possibles d'extraction des objets de même niveau : stratégie 1=modification des formulaires (niveau de dépendance et rapatriement des résultats d'interrogation sous forme textuelle balisée), stratégie 2=modification des formulaires pour le rapatriement des résultats de l'interrogation sous forme textuelle balisée et mise en place de procédures d'interrogation dans le langage de requêtes de STELLA ou d'XML, stratégie 3=extraction des fragments du TLFi correspondant à l'ensemble des entrées représentatives sous forme textuelle balisée et mise en place de procédures d'interrogation dans le langage de requêtes de STELLA ou d'XML.

La seconde étape consistera en **l'analyse des objets** de même niveau et même ancêtre via l'analyseur de G. Reb et N. Louis. Une comparaison pourra être faite en utilisant d'autres analyseurs du français en fonction des possibilités offertes par les participants au projet.

La troisième étape consistera en l'élaboration d'un moteur d'inférences permettant de **comparer le degré de spécificité des résultats de l'analyse** et de construire une ontologie basée sur la relation de subsomption. Dans le cadre de cette étape aussi, sera mise en œuvre la traduction des résultats de l'analyse dans le formalisme des logiques de description, suivie de la comparaison des résultats via le moteur d'inférence RACER.

A l'issue de cette troisième étape, un manuel de référence sera élaboré pour décrire en détail les procédures d'extraction d'informations sémantiques à partir des entrées du TLFi. Cela donnera lieu aussi à une nouvelle réunion plénière dont l'objectif sera de familiariser l'ensemble des participants au projet à l'utilisation et la mise en œuvre dans leurs équipes respectives des procédures d'extraction.

La quatrième et dernière étape sera centrée sur **l'exploitation des résultats** obtenus à ce stade. Les axes de perspective seront mis en œuvre : l'axe de la modélisation, celui de la validation d'hypothèses, celui de la constitution de lexiques sémantiques pour la résolution de la référence ou pour la génération d'inférences et celui de la construction d'un réseau sémantique suffisamment riche pour permettre l'élaboration de dictionnaires coopératifs.

La dernière réunion pleinière du projet se déroulera sous la forme d'une conférence ouverte au grand public où les membres ou d'autres scientifiques concernés présenteront les résultats du projet ou bien une approche critique des résultats de ce projet.

7. Demande de financement ILF ; autres sources de financement

(indiquer durée et budget prévisibles du projet, indiquer type de dépenses : vacations, missions, matériel, etc. envisagés)

Le budget demandé est de l'ordre de 30 K euros sur une durée de deux ans. Ce budget se répartit de la manière suivante :

- ✓ **Equipement** : 10K euros. L'essentiel de ce poste sera consacré à la mise en place de deux serveurs spécifiques pour l'accueil des données et des résultats du projet. L'un deux au moins sera implanté à l'ATILF mais accessible via une procédure sécurisée par tous les membres du projet. Compte-tenu de l'importance des ressources sous-jacentes et la nécessité d'un accès sécurisé aux données XML du TLF, l'accès aux données utiles pour le projet ne peut être réalisé, ni par une diffusion des données dans les différents laboratoires impliqués, ni par une ouverture totale de l'interface actuelle du TLFi.
- ✓ **Missions** : 6,7 K euros. Cette somme se répartit entre deux réunions plénières et plusieurs réunions restreintes. Les réunions plénières (4700 euros) sont a priori prévues à Nancy et prévoient le financement complet de deux intervenants de Nice (2*1100 euros), deux intervenants de Toulouse (2*1100 euros) et deux intervenants de Paris (2*300 euros). Le budget des réunions restreintes, à définir avec les membres du projet et en fonction des rencontres nécessaires, est d'environ 2000 euros.
- ✓ **Vacations** : 9 K euros. Sachant que plusieurs phases de codage seront nécessaires, nous prévoyons a priori le financement de six mois de vacations. A charge ensuite pour les membres du projet de décider où se dérouleront ces phases de codage.
- ✓ **Fonctionnement** : 4,3 K euros. Ce budget sera divisé entre la gestion du projet, à hauteur de 1300 euros, et l'organisation d'un workshop (financement complet des membres du projet sur le mode des réunions plénières et organisation du workshop), à hauteur de 3000 euros.