Les ressources de l'ATILF pour l'analyse lexicale et textuelle : TLFi, Frantext et le logiciel Stella

Pascale Bernard, Jacques Dendien, Josette Lecomte, Jean-Marie Pierrel.

ATILF-CNRS (Analyses et Traitements Informatiques du Lexique Français, UMR 7118) – 44, Avenue de la Libération – BP30687 – F-54063 Nancy-Cedex – France

pascale.bernard@inalf.fr, jacques.dendien@inalf.fr, josette.lecomte@inalf.fr, jean-marie.pierrel@inalf.fr

Abstract

Progress in the area of Linguistic Research, particularly in the field of Automatic and Statistical Treatment of Textual Data, depends on the accessibility to a vast amount of linguistic resources: text, lexicons, softwares. In order not to exclude anyone of the process of distributing these tools, it seems interesting to propose a mutualization of these resources to the benefit of the entire community. This paper presents the computerized linguistic resources of the Research Laboratory ATILF (Analyses et Traitements Informatiques du Lexique Français) available via the Web, and the wide range of their possible uses and applications. The Research Laboratory ATILF is the new UMR (Unité Mixte de Recherche) created in association between the CNRS and the University of Nancy 2 since January 2nd, and succeeds to the local component of the INaLF situated in Nancy. This considerable amount of resources concerning French language consists in a set of more than 3300 literary works grouped together in Frantext, plus a number of dictionaries, lexis and other databases. These web available resources are operated and run through the potentialities and powerful capacities of a software called Stella, a search engine specially dedicated to textual databases and relying on a new theory of textual objects. The general policy of our laboratory is to welcome and give the research and teaching world the widest access to all our resources.

Résumé

Les recherches en traitements automatiques et statistiques de données textuelles nécessitent pour progresser de vastes ressources : corpus textuels, dictionnaires informatiques, outils de traitement, souvent inaccessibles pour une équipe seule et qu'il convient donc de mutualiser au sein de la communauté de recherche du domaine. Cette contribution présente les ressources linguistiques informatisées du laboratoire ATILF (Analyses et Traitements Informatiques du Lexique français) disponibles sur la toile et leur diversité d'exploitation potentielle. L'ATILF est la nouvelle UMR créée en association entre le CNRS et l'Université Nancy 2 qui, depuis le 2 janvier 2001, a succédé à la composante nancéienne de l'INaLF. Ces importantes ressources sur la langue française regroupent un ensemble de plus de 3300 textes réunis dans Frantext et divers dictionnaires, lexiques et autres bases de données. Ces ressources exploitent les fonctionnalités du logiciel Stella, qui correspond à un véritable moteur de recherche dédié aux bases textuelles s'appuyant sur une nouvelle théorie des objets textuels. La politique du laboratoire consiste à ouvrir très largement ses ressources en particulier au monde de la recherche et de l'enseignement.

Mots-clés : bases textuelles, bases de connaissances, bases lexicographiques, langue française

1. Introduction

L'ingénierie des langues est devenue, au cours des dernières années, un des domaines-clés pour répondre aux besoins de notre société en terme d'analyse et exploitation de gisements d'information, le plus souvent sous forme textuelle, aujourd'hui disponibles (Pierrel, 2000).

Une rapide analyse de l'évolution de la linguistique au cours du dernier demi-siècle montre que sa confrontation avec l'informatique et les mathématiques a permis à la linguistique de se définir de nouvelles approches. C'est ainsi qu'au-delà d'une simple linguistique descriptive s'est développée une *linguistique formelle*, couvrant aussi bien les aspects lexicaux que syntaxiques ou sémantiques, qui tend à proposer des modèles s'appuyant sur une double validation, *explicative* d'un point de vue linguistique, *opératoire* d'un point de vue informatique. C'est elle aussi qui a permis l'émergence d'une véritable *linguistique de corpus* (Habert et al., 1998) permettant au linguiste d'aller au-delà de l'accumulation de faits de langue et de confronter ses théories à l'usage effectif de la langue. Parallèlement, les besoins applicatifs ont conduit à de nombreux travaux en informatique en *traitement automatique des langues*. Aujourd'hui traitement automatique des langues et linguistique de corpus tendent à structurer un nouveau champ disciplinaire dont les finalités sont multiples, en particulier :

- tout d'abord la mise en place d'applications concrètes : indexation et accès à l'information, résumé de textes, extraction de connaissances, dialogue homme-machine, par exemple ;
- la modélisation de la langue, de sa structure et de son usage qui conduit la linguistique, qui pendant longtemps demeura descriptive, à des exigences d'opérationnalité effective sur les formes d'usage effectif de la langue, par opposition aux exemples construits trop souvent encore utilisés en linguistique.

Ces études et recherches en traitement automatique des langues et en linguistique de corpus nécessitent de plus en plus l'usage de vastes ressources linguistiques : textes et corpus, si possible annotés (Véronis, 2000), dictionnaires et outils de gestion et d'analyse de ces ressources. Le coût de réalisation de telles ressources justifie pleinement des efforts de normalisation (Bonhomme, 2000) et de mutualisation pour permettre à la communauté de recherche de bénéficier pour le français de ressources comparables à celles existantes pour d'autres grandes langues telle l'anglais. Dans cet article nous avons choisi de présenter les ressources (bases de données textuelles, dictionnaires et outils d'exploitation) mises au point au sein de notre laboratoire héritier de l'INaLF (Institut National de la Langue Française) dont les deux caractéristiques premières sont, d'une part, leur qualité proprement linguistique, d'autre part, leur disponibilité via le web à l'adresse : http://www.inalf.fr/atilf/

2. Un ensemble intégré de ressources pour l'analyse lexicale et textuelle

Le laboratoire ATILF offre donc un ensemble de ressources informatisées (Bernard et al., 2001) composé de bases textuelles et de bases lexicologiques pour l'analyse lexicale et textuelle. Dans cet ensemble de ressources on distingue essentiellement la base Frantext, le Trésor de la langue française informatisé et sa version informatisée TLFi (CNRS, 1976-1994), et les 8° (1835) et 9° (en cours de rédaction) éditions du dictionnaire de l'Académie française. Une nouvelle fonction d'hypernavigation entre ces ressources a été récemment développée au laboratoire permettant une consultation croisée de tous ces produits à l'aide du logiciel Stella.

2.1. Dictionnaires : TLFi http://www.inalf.fr/tlfi/

2.1.1. Présentation générale :

Le Trésor de la langue française informatisé (Dendien, 1996) se présente à la fois comme une base lexicologique et une base de connaissances dont l'accessibilité est immédiate via l'internet. Le TLFi se distingue des autres dictionnaires électroniques existants, par la finesse de la structuration des données en « objets » interrogeables selon divers critères, et par une interface simple et conviviale qui offre trois niveaux de consultation via le logiciel STELLA.

Ces trois niveaux de requêtes correspondent à des besoins différents des utilisateurs. Il est possible de :

- simplement lire le dictionnaire, article par article, en mettant en évidence ou non tel ou tel type d'information (définition, exemple…) ;
- consulter le dictionnaire de façon transversale, par une requête élémentaire utilisant certains critères (définitions dans le domaine de la cuisine…) ;
- consulter le dictionnaire par une requête plus complexe croisant plusieurs critères. Ces requêtes peuvent être élémentaires ou multi-objets (on peut par exemple extraire tous les mots se terminant par le suffixe –âtre et extraire de cette liste ceux qui ont un sens péjoratif). Elles peuvent inclure ou exclure un contenu, etc..

Il possède une fonctionnalité supplémentaire : l'hypernavigabilité et ses liens avec d'autres produits ATILF : Frantext (textes du domaine public uniquement), et les dictionnaires de l'Académie française.

2.1.2. Spécificités en terme de contenu :

Le TLFi se distingue essentiellement par la richesse de son matériau et la complexité de sa structure :

- Originalité de sa nomenclature, les mots sont présents dans nos fonds ou dans des dictionnaires ; originalité aussi par le traitement de certains morphèmes (plus de 60 mots traités sous le morphème *–o*), et par le traitement des préfixes, suffixes et autres éléments formants : c'est en tout 100 000 mots avec leur étymologie et leur histoire, 270 000 définitions.
- Richesse des objets métatextuels (vedettes, codes grammaticaux, indicateurs sémantiques ou stylistiques, indicateurs de domaines, définitions, exemples référencés…). En tout, plus d'une quarantaine d'objets différents.
- Richesse des 430 000 exemples, tirés de deux siècles de littérature française.
- Diversité des rubriques : une rubrique de synchronie couvrant la période 1789 à nos jours ; une rubrique d'étymologie et d'histoire, et une rubrique de bibliographie termine les principaux les articles.

2.1.3. Spécificités en terme de balisage :

Un des principaux avantages d'un dictionnaire informatisé est de permettre d'effectuer des recherches transversales "plein texte" à travers la totalité de son contenu. Cependant, chaque occurrence du texte cherché a une signification qui dépend essentiellement du **type** de l'objet textuel dans lequel elle est localisée. Restreindre une recherche "plein texte" à tel ou tel type donné permet donc de diminuer le bruit et d'accroître la précision des recherches. Afin de rendre les interrogations du TLF plus précises et significatives, il a été procédé à un balisage textuel XML de **tout le texte** du dictionnaire en y injectant des balises repérant le début, la fin et le type de chaque objet textuel rencontré. Une quarantaine de types d'objets différents ont ainsi été introduits à l'aide d'automates experts, alors que bien des dictionnaires informatisés s'arrêtent à quelques types essentiels (souvent définitions et/ou citations). Il est ainsi possible de limiter une recherche "plein texte" à l'un de ces types.

Afin d'introduire encore plus de précision dans les requêtes, il convenait d'ouvrir une nouvelle dimension : celle de la structure hiérarchique de chaque article. En effet un article de dictionnaire (à l'exception des articles les plus élémentaires) introduit une structure explicitée dans le TLF (comme dans bien d'autres dictionnaires) par des sigles de structure hiérarchisés

(I, II, …; A, B,…; 1, 2, …, a, b,…). Une indication de domaine technique "mécanique" apparaissant au niveau B, par exemple, signifie clairement que les éventuelles subdivisions hiérarchiques du B (et y compris les éventuelles subdivisions de ces subdivisions) traitent de la mécanique. Ainsi, une définition trouvée dans le paragraphe b) appartenant au paragraphe 2) qui appartient au B introduit nécessairement un sens usité dans le domaine de la mécanique. De manière générale, il est donc possible d'introduire systématiquement une relation entre deux objets X et Y : X est hiérarchiquement inférieur, égal ou supérieur à Y.

Il devient alors possible d'effectuer des requêtes comportant N objets, chaque objet ayant un contenu textuel imposé, et les liens relationnels entre objets étant imposés. Pour que la requête ait un sens, il suffit que le graphe dont les sommets sont les objets, et dont les arcs sont les relations hiérarchiques imposées soit connexe. Par exemple, soit une requête spécifiant qu'un objet A de type "catégorie grammaticale" contienne le mot "verbe", qu'un objet B de type "indication de domaine technique" contienne le mot marine, qu'un objet C de type "définition" contienne le mot "voile" ou "voiles", que A soit hiérarchiquement supérieur à B (ce qui signifie que l'indication "marine" est afférente à un verbe), et enfin que B soit hiérarchiquement supérieur à C (ce qui signifie que la définition est trouvée dans une section d'article traitant de marine). Une telle requête revient de toute évidence à rechercher tous les verbes utilisés dans la marine pour la manœuvre des voiles.

Afin de pouvoir gérer les interdépendances hiérarchiques entre objets, une deuxième série d'automates a été développée, capables d'analyser la structure hiérarchique des articles. Le résultat de ces automates est d'introduire un jeu complémentaire de balises textuelles rendant compte de la structure. La cohérence du système de balises est contrôlée par une grammaire (correspondant à une DTD XML) permettant une validation formelle de la structure.

En conclusion, il a été constaté que la richesse du balisage du TLF qui permet l'identification de nombreux types d'objets et leur mise en relation hiérarchique, contribue à obtenir des résultats ayant un degré de pertinence très élevé, chaque contrainte hiérarchique contribuant à filtrer le bruit. L'interface utilisateur permet d'exprimer de telles requêtes avec la plus grande facilité.

2.2. Bases textuelles : FRANTEXT http://www.inalf.fr/frantext/

2.2.1. Présentation générale:

Frantext (1992) peut se définir comme un doublet constitué d'une part d'un vaste corpus des textes de langue française et d'autre part d'un logiciel offrant une interface Web permettant son interrogation et sa consultation.

Historiquement, le but premier du corpus textuel (dont la saisie a débuté dès les années 60) était de permettre la constitution de "dossiers de mots" destinés aux rédacteurs du dictionnaire TLF. Un rédacteur ayant à écrire l'article du TLF consacré au mot "briquet", par exemple, se trouvait ainsi doté d'une concordance systématique de ce mot triée suivant différents critères (tri alphabétique ou par catégories grammaticales des contextes gauches ou droits). D'autre part le corpus textuel servait également, en phase finale de rédaction, à sélectionner le texte des exemples fournis dans le TLF. On peut ainsi estimer que ce corpus textuel a fourni environ 90% des quelque 430000 exemples cités dans le TLF.

La constitution des dossiers de mots, ainsi que les extractions des exemples finalement retenus, étaient assurées par de lourds logiciels non interactifs (de type traitement par lot) procédant par traitement séquentiel du corpus. Vers les années 80, le laboratoire a réalisé une plate-forme de type base de données textuelles qui a permis un gain de productivité spectaculaire grâce à la possibilité d'accès direct aux mots du corpus. Surtout, cette plate-forme a permis d'envisager la

réalisation d'une première interface utilisateur (1985), avec une exploitation télématique par les moyens de l'époque (terminaux Transpac, Minitel).

Progressivement, la raison d'être initiale de Frantext (au service du TLF) a été supplantée par le souci de mettre à la disposition de la communauté scientifique un corpus textuel de plus en plus élaboré doté d'un outil d'interrogation de plus en plus efficace.

2.2.2. Etat actuel du corpus

Le corpus actuel comporte 3350 textes dont les dates s'échelonnent de 1505 à 1997. Le corpus textuel de Frantext vient ainsi prolonger un corpus de français médiéval et de moyen français (environ 300 textes couvrant les années 847 à 1502 et dont la mise à disposition de la communauté scientifique est envisagée), offrant ainsi une des plus grandes ressources disponibles sur toutes les époques de la langue française. Il devrait être enrichi dans l'année à venir par un ensemble important de textes scientifiques contemporains.

Le corpus de Frantext contient environ 80% de textes littéraires (texte intégral) et 20% de textes techniques représentatifs des principales disciplines scientifiques.

Une veille permanente est assurée sur ce corpus avec plusieurs objectifs :

- Qualité de la saisie : une partie du corpus actuel est encore dans l'état de la saisie initiale (début des années 60). Ces textes ont quelquefois été victimes d'avatars liés à la technologie de l'époque (utilisation du très peu fiable et non regretté ruban perforé). Cette partie du corpus fait l'objet d'une campagne intensive de correction.
- Qualité des éditions saisies : une des originalités de Frantext est de se référer à des éditions dûment référencées (on regrettera à cet égard qu'il en soit rarement de même pour les nombreuses sources textuelles communément trouvées sur Internet), condition sine qua non pour qu'une citation puisse avoir de la valeur. Certaines de ces éditions peuvent être considérées comme obsolètes par les spécialistes et, en ce cas, des éditions plus récentes leur sont de plus en plus substituées.
- Complémentation du corpus : bien entendu, on pourra toujours regretter l'absence de tel ou tel texte dans le corpus de Frantext, mais nul ne peut échapper à cette critique. Il existe cependant une politique de complémentation du corpus axée sur quelques principes stables : d'une part, rééquilibrer les différentes époques et les différents genres, d'autre part faciliter des opérations précises de recherche ou d'enseignement telle la mise à disposition du public de la totalité des textes inscrits au programme de l'agrégation 2002.

2.2.3. Les deux Frantext

Frantext est disponible sur l'internet, moyennant abonnement de 305 €, à ce jour sous deux formes :

a) La totalité du corpus (3350 textes) que l'on peut interroger sur les formes graphiques du texte.

b) La partie du corpus en orthographe "moderne" (1940 textes) entièrement étiqueté en catégories grammaticales par un logiciel de catégorisation réalisé par l'ATILF. Cette version de Frantext peut être interrogée à la fois sur les formes graphiques et sur les catégories grammaticales.

3. Le logiciel STELLA : boîte à outils d'exploitation de ressources textuelles

3.1. Présentation générale

Stella est le logiciel qui anime la base de données Frantext ainsi que le TLFi. Il a été intégralement développé à l'ATILF. Ce logiciel peut s'appliquer à tout ensemble de données textuelles, structurées ou non. Outre Frantext et le TLFi, on trouvera sur Internet des versions informatisées du dictionnaire de l'Académie française : 8e édition de 1935 http://zeus.inalf.fr/academie.htm et 9e édition (en cours) http://zeus.inalf.fr/academie9.htm.

Le logiciel Stella se présente comme une boîte à outils (C++) comportant différents volets :

a) Utilitaires divers incluant tris, traitement des expressions régulières, et surtout une base de données fondée sur la nomenclature du TLF permettant des opérations de flexion ou de lemmatisation.

b) Interface Web permettant la mise en œuvre facile d'interfaces utilisateur, des fonctions de gestion de " sessions utilisateur " et de gestion d'espace de travail sur le serveur ainsi qu'une solution permettant une hypernavigation entre les différentes applications gérées par Stella, qu'elles résident ou non sur un même serveur. L'hypernavigation permet une liaison dynamique entre les différentes ressources textuelles : il est ainsi possible, en cliquant sur n'importe quel mot d'une page affichée par l'une des bases gérées par Stella de déclencher l'apparition d'un menu déroulant proposant de le rechercher dans n'importe laquelle des autres bases. Par exemple, le fait de cliquer sur le mot " aimerions " permet d'afficher l'article " aimer " du TLFi ou de la 8e ou la 9e édition de l'Académie.

c) Un système complet de gestion de base textuelle assurant à la fois les fonctions de stockage et d'accès à l'information.

3.2. Les spécificités de Stella

La conception de Stella, en tant que logiciel de gestion de base textuelle, repose sur des principes mathématiques rigoureux (Dendien, 1991) :

3.2.1. Fonctions de " bas niveau "

Ces fonctions permettent de fabriquer et de gérer les structures de stockage de l'information. Elles reposent sur un système d'indexation garantissant un stockage optimal des données en terme d'encombrement. Il est mathématiquement démontré que ce système a des performances égales à celles qui sont prévisibles par la Théorie de l'Information, avec la propriété remarquable que la quantité d'information nécessaire pour coder l'emplacement d'une occurrence d'un mot est indépendante de la taille du corpus (elle ne dépend que de la probabilité de ce mot), ce qui rend ce système apte à gérer des corpus de taille considérable.

Qui plus est, la compacité optimale du système d'indexation assure des performances optimales en minimisant les échanges entre mémoire de masse et mémoire centrale.

3.2.2. Fonction de " haut niveau "

De même que les fonctions de " bas niveau " permettant une localisation rapide des occurrences d'un objet textuel élémentaire (mot) au sein du corpus, les fonctions de haut niveau permettent de s'attaquer au même problème pour la localisation d'objets textuels beaucoup plus complexes. Ces fonctions de haut niveau reposent sur une théorie des objets textuels dont les principes sont les suivants :

- Tout objet textuel est manipulable par des méthodes standard (dans l'implémentation informatique, il s'agit de méthodes virtuelles d'une classe de base) permettant d'implémenter son moteur de recherche.
- Un ensemble de lois de composition (dont les plus simples sont les listes et les séquences) permet de fabriquer des objets textuels nouveaux, dits objets **composites**, à l'aide d'objets élémentaires ou d'autres objets composites. Les méthodes standard de manipulation d'un objet composite se déduisent, en fonction de la loi de composition appliquée, des méthodes des objets ayant servi à le fabriquer.

De ces principes, on peut déduire les conséquences suivantes :

- Il est possible de fabriquer des objets composites d'une complexité arbitrairement élevée en appliquant successivement les lois de composition autant de fois que l'on veut.
- Tout objet composite est automatiquement muni de son moteur de recherche, qui se déduit des moteurs de recherche de ses composants, et, par récursivité, des moteurs de recherche des objets élémentaires à l'origine de sa construction.
- Il est possible de fabriquer des objets élémentaires nouveaux, dits objets **natifs**, de manière tout à fait arbitraire : il suffit pour cela de les doter des méthodes standard. Ces objets natifs pourront à leur tour rentrer comme éléments de fabrication d'objets composites. Par exemple, si on définit un objet " joker " muni de méthodes de localisation (particulièrement faciles à implémenter car elles expriment que ce joker est un objet situé n'importe où), il devient possible de définir un objet composite de type " séquence " comportant successivement le mot " un ", le joker, et le mot " homme " pour localiser toutes les occurrences de " un XX homme ", dans laquelle XX désigne un mot quelconque. Un autre exemple de constructeur d'objet natif admettant comme paramètre l'infinitif d'un verbe, consiste à fabriquer la liste composée des objets élémentaires correspondant à chacune de ses formes fléchies. Le moteur de recherche de cet objet localisera l'ensemble de formes fléchies du verbe donné en paramètre.

Ce mécanisme de création d'objets composites ou natifs confère à Stella une **architecture totalement ouverte** à bien des égards :

- L'ensemble des lois de composition est ouvert : l'implémentation de lois nouvelles permet la création d'objets composites nouveaux. Par exemple, si un texte a été segmenté et étiqueté à l'aide d'un analyseur morpho-syntaxique, il est possible de créer (comme cela est fait dans Frantext) un objet composite représentant à la fois le contenu textuel et les attributs grammaticaux associés (par ex. contenu " tire-bouchon " et attribut " substantif).
- L'ensemble des objets natifs peut s'enrichir à l'infini. Tout nouvel objet natif peut entrer dans la fabrication d'objets composites, démultipliant ainsi leur combinatoire.
- Les objets natifs nouveaux (on l'a vu plus haut dans l'implémentation des formes fléchies d'un verbe) permettent d'introduire facilement un " savoir linguistique ". Par exemple, rien n'est plus facile que d'introduire, si on le désire, les déclinaisons latines.

3.2.3. Les grammaires

La possibilité de fabriquer des objets arbitrairement complexes a pour seule limite la puissance de description de ces objets offerte dans l'interface utilisateur. Dans ce domaine, Frantext va bien au-delà de ce qui est habituellement offert par les autres systèmes d'expression de requêtes en permettant à l'utilisateur d'exprimer ses requêtes grâce à des répertoires de règles réutilisables et paramétrables appelés grammaires. Pour aborder cet aspect, nous proposons

d'examiner deux exemples : l'étude des utilisations pronominales d'un verbe donné et l'étude des énumérations.

3.3. Exemples d'utilisation de grammaires dans Stella

3.3.1. Usages pronominaux d'un verbe

Si on se propose de rechercher, dans un corpus donné, les utilisations pronominales d'un verbe donné, on se trouve confronté à une difficulté due au caractère multiforme de ces utilisations (tournures affirmatives, interrogatives, négatives, interro-négatives, temps simples ou composés). Cette difficulté rend une telle recherche illusoire avec la quasi-totalité des systèmes existants. On trouvera ci-dessous un exemple (simplifié !) de grammaire permettant de détecter la plupart des occurrences des usages pronominaux du verbe " laver ", hormis les tournures interrogatives. Les commentaires sont spécifiés en italique entre crochets, ex : [*commentaire*] .

Dans la grammaire qui suit, les lignes **en gras** correspondent à des déclarations de règles de grammaire. Une règle XXX peut être invoquée dans une autre règle par la syntaxe &rXXX. Toute règle invoquée doit être déclarée (que ce soit en amont ou en aval de son invocation)

[*Règle décrivant le discours trouvé devant le verbe dans les tournures affirmatives*.]
preambule_affirmatif :
je (me|m') | tu(te|t') | (se|s') | nous nous | vous vous
[*Régle décrivant une tournure affirmative à un temps simple (&claver désigne une forme fléchie du verbe " laver ")*]
temps_simple_affirmatif :
&rpreambule_affirmatif &claver
[*Idem pour une tournure affirmative à un temps composé . &cêtre désigne une forme fléchie du verbe " être "*]
temps_compose_affirmatif :
&rpreambule_affirmatif &cêtre &rparticipe_passe
participe_passe :
lavé | lavée | lavés | lavées
[*Règle décrivant le discours trouvé devant le verbe dans les tournures négatives*.]
preambule_negatif :
je ne (me|m') | tu ne (te|t') | ne (se|s') | nous ne nous | vous ne vous
[*Description d'une tournure négative, temps simple*.]
temps_simple_negatif :
&rpreambule_negatif &claver &rfin_negation
[*Description d'une tournure négative, temps simple*.]
temps_compose_negatif :
&rpreambule_negatif &cêtre &rfin_negation &rparticipe_passe
[*Seconds termes possibles d'une négation*]
fin_negation :
pas|plus|jamais|guère|mie|point
[*La règle usage_pronominal ci-dessous réunit les différents cas*.]
usage_pronominal :
&rtemps_simple_affirmatif | &rtemps_compose_affirmatif | &rtemps_simple_negatif | &rtemps_compose_negatif

Cette grammaire peut être utilisée dans une requête en invoquant une de ses règles :

- &rtemps_simple_negatif invoque la règle permettant de localiser les usages pronominaux à un temps simple dans une tournure négative.
- &rusage_pronominal invoque la règle permettant de localiser tous les usages pronominaux.

Voici à titre d'exemple un extrait des résultats obtenus par l'application de cette grammaire sur un sous corpus de Frantext, donnant ainsi un ensemble diversifié d'exemples attestés dans la littérature (Frantext fournit directement la référence de chaque exemple) :

a) Temps simple (forme affirmative)

... exigeait que *nous nous lavions* les mains en même temps que lui.
Rouaud J. / Les champs d'honneur

…personne ne s'emploie à enseigner à un enfant à *se laver* tout seul.
Dolto F. / La cause des enfants

b) Temps simple (forme négative)

…car *ils ne se laveront jamais* de la honte dont ils se sont couverts…
Balzac H. / Le médecin de campagne

Il parlait peu, d'un ton bourru, et *ne se lavait pas* davantage.
Caradec F. / La compagnie des zincs

..cela doit faire huit jours que *je ne me lave plus*.
Duras M. / La douleur

c) Temps composé (forme affirmative)

Que *je m'étais lavé* les pieds en vain.
Brassens G. / Poèmes et chansons

... puis quand *vous vous êtes lavé* le visage à l'eau de roses…
Giraudoux J. / La folle de Chaillot

d) Temps composé (tournure négative)

…et qui *ne s'était pas lavé* les mains pour les faire paraître calleuses.
Flaubert G. / L'éducation sentimentale

…elles qui *ne s'étaient jamais lavées* que dans des éviers…
Simon C. / L'acacia

Bien entendu, cette grammaire peut être complétée pour les tournures interrogatives et interro-négatives. Il est également possible, bien entendu, de la rendre paramétrable pour fonctionner avec un verbe quelconque.

3.3.2. Etude d'énumérations (dans Frantext catégorisé)

Fréquemment, certains auteurs procèdent à des énumérations (veau, vache, cochon, couvée ; à pied, à cheval, en voiture, etc.). Voici une proposition pour aborder le problème de la détection avec une grammaire de Stella :

*[La règle **"item"** définit l'item textuel qui va se répéter. Elle admet deux paramètres &1 et &2 qui seront remplacés par leurs valeurs effectives correspondante lors de l'invocation de la règle. Par exemple:*
1) &ritem(en,S) invoque la règle item en passant "en" et "S" en paramètre. La règle item devient alors équivalente à "en &e(g=S)" ce qui signifie "en" suivi d'un substantif.
2) &ritem(,A) invoque la règle item en passant un premier paramètre vide et "A" en second paramètre. La règle item devient alors équivalente à "&e(g=A)" ce qui désigne un adjectif.]

item :
&1 &e(g=&2)

[*La règle **"repetition"** exprime que l'item textuel doit se répéter deux fois (sous-expression "&ritem(&1,&2) , &ritem(&1,&2)" plus un nombre (supérieur ou égal à 1) quelconque de fois (sous expression "&+(, &ritem(&1,&2))"). L'item doit donc se répéter 3 fois ou plus.*
Cette règle permet donc de rechercher un item textuel se répétant au moins 3 fois avec une virgule intercalaire.]

repetition :
&ritem(&1,&2) , &ritem(&1,&2) &+(, &ritem(&1,&2))

Pour utiliser cette grammaire (nommée par ex. G1) il suffit d'invoquer la règle "repetition" en lui passant les deux paramètres qu'elle transmettra à la règle "item". Voici à titre d'exemple quelques résultats fournis par cette grammaire sur des textes de Zola :

> &rrepetition(,A),G1 détecte : ***"beau, jeune, amoureux, écervelé"*** (La curée); ***"liberticide, anticonstitutionnel, révolutionnaire"*** (Le ventre de Paris) ; ***"rouges, jaunes, mauves, blanches"*** (La faute de l'abbé Mouret), etc.
>
> &rrepetition(en,S),G1 détecte : ***"en rumeurs, en rires, en applaudissements"*** (Nana) ; ***"en coeurs, en festons, en médaillons"*** (L'œuvre) **;** ***"en vertu, en génie, en crime, en ivrognerie, en sainteté"*** (Le docteur Pascal), etc.
>
> &rrepetition(,V),G1 détecte : "***chancelait, glissait, tombait***" (La conquête de Plassans) **;** ***"débattait, mordait, égratignait"*** **(**La joie de vivre) **;** ***"rampait, descendait, remontait"*** (Germinal), etc.

Ces deux exemples de grammaires tendent à prouver que tous les linguistes, syntacticiens aussi bien que sémanticiens, stylisticiens et autres peuvent exploiter avec bonheur les possibilités offertes par Stella sur les ressources offertes par l'ATILF.

4. Diversité d'exploitations possibles

Outre cette possibilité d'écriture et d'exploitation de grammaires, Frantext et le TLFi offrent de nombreux services qu'il serait fastidieux d'énumérer et de décrire ici. En un mot les possibilités d'interrogation s'articulent autour des axes suivants :

4.1. Recherche de cooccurrences et collocations:

La taille du corpus Frantext fait que l'on dispose d'une couverture linguistique importante.

Frantext offre différents services, notamment l'étude du vocabulaire au voisinage des occurrences d'un mot donné, ce qui est très utile dans des études thématiques ou des recherches de collocations. Les résultats rendus par ce service sont constitués de la liste des mots trouvés au voisinage du mot donné, triés au choix par ordre alphabétique, ordre croissant ou décroissant des fréquences.

Dans le TLFi : il est possible de rechercher également les collocations, les locutions si on est intéressé par un travail sur la phraséologie.

4.2. Extraction de sous-lexiques

Dans Frantext : le logiciel permet de constituer un sous corpus en sélectionnant à l'intérieur de la base un ensemble de textes sur des critères variés pouvant se combiner entre eux : auteurs, dates, périodes ou genres.

Dans le TLFi : le logiciel permet une interrogation par domaines ou tout autre type d'objets : code grammatical, indicateurs sémantiques ou stylistiques…. Par exemple, on peut y rechercher des graphies particulières, mais aussi des proverbes, des locutions. On peut aussi créer des lexiques de domaines techniques particuliers (la marine, mais aussi la mythologie ou l'œnologie) pour peu que les indicateurs correspondants figurent dans les articles du TLF papier.

4.3. Etudes morphologiques

Dans Frantext, on peut vouloir rechercher des exemples de dérivation ou de composition de mots. Par exemple, les éléments *homme-* ou *femme-* en composition, pour les comparer à des séquences moins figées (sans traits d'union), tant au plan de leur environnement contextuel que

de leurs fréquences d'attestation dans la littérature. On peut aussi, par exemple exploiter des listes de fréquence concernant certaines formes verbales : Est-ce que *assied* est plus fréquemment utilisé que *assoit* ?

Dans Frantext catégorisé, comme dans le TLFi, on peut rechercher la liste des verbes en –er, mais avec des visées différentes. Dans le TLFi, on pourra coupler la requête avec la notion de transitivité par exemple : tous les verbes en –er qui sont indiqués transitifs (au niveau du code ou au niveau d'un indicateur). On pourra aussi s'intéresser aux rubriques Etymologie et Histoire de ce mot, et/ou à toutes les Remarques associées à l'emploi de ce mot, par une visualisation simple de l'article.

La consultation de ces deux outils permettrait d'éviter d'avancer brutalement des théories non étayées par les faits ou, plus positivement, de suggérer des hypothèses nouvelles sur des faits de langue bien attestés.

4.4. Etudes de Syntaxe locale

Pour repérer des motifs récurrents ou des différences syntaxiques ou autres phénomènes linguistiques pertinents, cet ensemble de ressources est d'une grande importance. Par exemple, il est possible de repérer les constructions différentes autour d'un même mot (travailler son style, travailler à sa thèse, travailler du chapeau,…). Il est de même possible, dans le TLFi, de rechercher les syntagmes contenant en début un verbe infinitif suivi de la préposition *en*. Ou bien encore tous les verbes du 1er groupe intransitifs (1221 solutions) pour étudier leur environnement à partir des exemples correspondants présents dans l'article.

La consultation systématique de Frantext et du TLFi est riche d'enseignements pour approfondir la connaissance du comportement de *comme* et sa classification selon les parties du discours. C'est un bon outil pour repérer les différences de positionnement des mots dans un syntagme (Adjectifs antéposés /postposés) ou leur combinatoire grammaticale.

4.5. Etudes de sémantique

Dans le TLFi, par exemple, il est possible de rechercher les mots contenant un suffixe *–esque* ou un préfixe *re-* pour repérer les éléments auxquels ils sont accrochés, mais aussi dans le but de définir plus précisément les sens de ces affixes et leur portée. Ou encore, on peut souhaiter trouver toutes les entrées dont la définition contient le mot « outil » en tout début de définition. Une recherche complexe amène 245 réponses. Une requête plus fine, sur les entrées en *–oir* dont la définition contient le mot « *outil* », sera un peu plus compliquée (Gestion de listes/Création à partir des graphies du TLF/critère= .*oir . Puis travail sur cette liste, pour y repérer une association d'un de ses éléments avec une définition contenant « outil »). Cette requête amène 54 réponses. On peut aussi exploiter le TLFi comme une riche base de synonymes…. On pourra trouver dans (Martin, 2001) divers autres exemples d'exploitation possible du TLFi pour des études en sémantique.

4.6. Etudes de Stylistique

Dans le TLFi, il est possible, par exemple, de demander tous les exemples de Colette contenant une forme conjuguée du verbe *aimer* dans le corps de l'exemple. Frantext est aussi un outil performant pour les études des vocabulaires spécifiques à un auteur.

Le TLF a pris en compte les néologismes, hapax (à la date de sa constitution). On peut les chercher dans le TLF informatisé, et ainsi avoir une idée de l'évolution de la langue en ce qui les concerne.

4.7. Autres possibilités

Frantext, TLFi et Dictionnaires de l'Académie informatisés sont des outils interconnectés, avec des liens d'hypernavigation. Ce qui ajoute une richesse supplémentaire.

Dans le TLFi, on peut retrouver les origines et premières attestations des mots, tant dans le domaine médical que philosophique ou technique ou autre. Les étymologistes peuvent donc être intéressés par sa consultation.

Il y existe des rubriques Prononciation qui peuvent intéresser des phonéticiens et phonologues. Certains articles, particulièrement ceux concernant les mots grammaticaux, sont riches de remarques concernant les indications d'emploi. De même ceux concernant les éléments dits « formants ».

Frantext devrait être un passage obligé pour toutes les études de fréquence concernant les mots du français.

5. Conclusion et perspectives

Frantext, TLFi et Dictionnaires de l'Académie informatisés forment un ensemble, mais ne sont pas fermés sur eux-mêmes. Cet ensemble de corpus, ouvert à toutes sortes d'utilisations dans tous les domaines, ne représente qu'une partie des richesses de l'ATILF (Bernard et al., 2000).

Cet ensemble de ressources brutes est mis à la disposition de la Communauté scientifique car il peut être le fondement de projets divers et variés, en particulier la constitution de lexiques dérivés, à partir d'études portant sur les synonymes, ou sur les collocations. Plusieurs projets de collaboration sont à l'étude, entre l'ATILF et des partenaires intéressés, visant de nouveaux objectifs scientifiques à partir des ressources offertes. Conscients que ces ressources constituent un patrimoine important financé essentiellement par le CNRS, nous avons le souhait de les mettre le plus possible à disposition de la communauté de recherche à travers des projets coopératifs de recherche. Toute nouvelle proposition de coopération est vivement souhaitée et sera étudiée avec soin.

Références

Bernard P., Bernet C., Dendien J, Pierrel J.M., Souvay G et Tucsnak Z (2001) « Un serveur de ressources linguistiques informatisées via le Web », *Actes de TALN 2001*, Tours, Juillet 2001, pages 333-338.

Bonhomme P. (2000) *Codage et normalisation de ressources textuelles*, in (Pierrel, 2000).

CNRS (1976-1994) *TLF, Dictionnaire de la langue du 19e et 20e siècle*, CNRS, Gallimard, Paris.

Dendien J. (1991) Access to information in a textual database: access functions and optimal indexes, Oxford, Clarendon press.

Dendien J. (1996). Le projet d'informatisation du TLF, in *Lexicographie et informatique*, pages 25-34.

FRANTEXT (1992) *Autour d'une base de données textuelles ; témoignages d'utilisateurs et voies nouvelles*, Paris, Didier Érudition.

Habert B., Nazarenko A., Salem A. (1997) *Les linguistiques de corpus*, Arman Colin, Paris.

Martin R. (2001) *Sémantique et automate*, Ecritures électroniques, PUF, Paris.

Pierrel J.M. (2000) *Ingénierie des Langues*, Traité Information - Commande - Communication, Editions Hermès, octobre 2000.

Véronis J. (2000) *Annotation automatique de corpus : panorama et état de la technique*, in (Pierrel, 2000).