La technologie au service de l'étude de la langue :corpus et dictionnaires informatisés.

Jean-Marie Pierrel

La technologie au service de l'étude de la langue : corpus et dictionnaires informatisés

Jean-Marie Pierrel

ATILF Université Henri Poincaré Nancy 1
44 Av De la Libération BP 30687 54063 Nancy Cédex
Jean-Marie.Pierrel@atilf.fr

Résumé - Abstract

L'évolution, au cours des dernières années, des technologies de l'information et de la communication modifie profondément les méthodes de recherche en sciences humaines et sociales. La notion de documents ou de corpus d'étude informatisés est en effet de plus en plus incontournable dans nos disciplines que cela soit en linguistique, en langue, en littérature, en histoire, en droit ou en économie. Nous pensons que le domaine du droit ne peut échapper à cette évolution. Nous avons choisi de présenter, dans cet article, les ressources (bases de données textuelles, dictionnaires et outils d'exploitation) mises au point dans notre laboratoire dont les caractéristiques de qualité proprement linguistique et de disponibilité en font des ressources de référence pour le français.

The evolution, during last years, of communication and information technologies deeply modifies the methods of research in social sciences. The concept of documents or computerized study corpora is indeed increasingly impossible to circumvent in our disciplines (linguistics, language, literature, history, law or economy). We think that the field of the law cannot escape from this evolution. We chose to present, in this article, the resources (textual data bases, dictionaries and tools) developed in our ATILF laboratory whose characteristics of properly linguistic quality and availability make of them French reference resources.

Keywords – Mots Clés

Computerized resources, corpus, dictionaries, TLFi, Frantext

Ressources informatisées, corpus, dictionnaires, TLFi, Frantext.

Introduction

L'évolution, au cours des dernières années, des technologies de l'information et de la communication et leur disponibilité actuelle modifient profondément les méthodes de recherche en sciences humaines et sociales. Les aspects d'exploitation scientifique de documents sont devenus particulièrement importants pour ne pas dire stratégiques. La notion de documents ou de corpus d'étude est en effet de plus en plus incontournable dans la plupart de nos disciplines que cela soit en linguistique, en langue, en littérature, en histoire, en droit ou en économie.

Aujourd'hui, un des aspects essentiels pour l'exploitation scientifique et la valorisation de documents à forte composante textuelle (ce qui n'exclut d'ailleurs pas des documents multimédia incluant des graphiques, des images ou du son) est leur informatisation et leur accès à travers des outils intelligents de recherche et d'étude qui ne se limitent pas à l'exploitation de simples mots clés ou d'informations décrivant leurs structures mais qui permettent un véritable accès par le contenu à travers soit une recherche plein texte, soit une exploitation d'annotations et de balisages représentatifs de leurs contenus informationnels (Véronis 2000).

La plupart des équipes de recherche de nos domaines, au-delà d'une approche scientifique classique, utilisent aujourd'hui de vastes ressources linguistiques – textes et corpus, si possible annotés, dictionnaires et lexiques spécialisés, outils de gestion et d'analyse de ces ressources – comme fondement d'une approche fortement renouvelée où la référence aux corpus est devenue un standard à la fois pour faire émerger des connaissances et leur modélisations et servir de base au processus nécessaire de validation et d'évaluation de ces résultats. Mais l'élaboration de ressources informatisées fiables entraîne un coût de réalisation qui justifie pleinement les efforts de normalisation (Bonhomme 2000) et de mutualisation qui ont vu le jour au cours de la dernière décennie pour permettre à l'ensemble de la communauté de recherche de bénéficier des ressources nécessaires à cette nouvelle approche.

Nous pensons que le domaine du droit, et plus particulièrement du droit comparé, ne peut échapper à cette évolution. Nous avons choisi de présenter, dans cet article, les ressources (bases de données textuelles, dictionnaires et outils d'exploitation) mises au point dans notre laboratoire dont les caractéristiques de qualité proprement linguistique, d'une part, et de disponibilité via le Web à l'adresse : www.atilf.fr, d'autre part, en font des ressources de référence pour le français.

Après avoir montré l'intérêt de disposer de ressources de références comme appui à nos recherches, nous présenterons plus particulièrement les ressources disponibles à l'ATILF, TLFi, Frantext et logiciel d'exploitation et, à travers quelques exemples, leurs intérêts d'usage dans le domaine de l'étude du droit.

2 Quelles ressources informatisées ?

1.0 Corpus textuels

Le premier type de ressources, indispensables pour le développement de nombreuses études en Sciences Humaines et Sociales, concerne les corpus textuels. Support de la connaissance accumulée aux cours des ans, leur rôle est en effet central pour permettre la construction de modèles avec comme objectif premier de proposer, parfaire et évaluer des modèles opératoires ou des théories représentatifs de l'usage effectif d'une notion dans tel ou tel domaine. Il s'agit par exemple de faire émerger des invariants ou, au contraire, des comportements particuliers d'entités linguistiques ou de lexèmes. Si, pendant longtemps, ce type d'activité a pu se satisfaire des connaissances intrinsèques du chercheur, les besoins de validation objective du monde scientifique ont progressivement imposé le recours à des corpus attestés. La question fondamentale est alors de savoir comment recueillir des données fiables sur les usages effectifs. Le Web peut être aujourd'hui une source importante d'extraction de connaissances, mais on peut à juste titre s'interroger sur la fiabilité des ressources textuelles qu'on y trouve ! Deux travers de taille caractérisent, en effet, les textes disponibles sur le Web :

1. Leur qualité est souvent très discutable. Sans parler des nombreuses erreurs qui demeurent dans bien des documents disponibles sur la toile, on y retrouve un mélange de textes de formes, de genres, de niveaux de langue et d'époques très disparates.

2. La pérennité de leur disponibilité n'est pas toujours assurée. Le propre du Web est de fournir des informations en constante évolution et leur durée de vie est souvent inférieure à la durée de vie des projets de recherche qu'elles sous-tendent. Or le principe même de l'évaluation scientifique et de la comparaison de travaux scientifiques nécessite d'assurer la possibilité de dupliquer des expérimentations scientifique d'évaluation sur corpus identiques.

La question de la qualité et de la disponibilité des corpus reste donc importante pour nos domaines de recherche et, pour s'en convaincre, il suffit d'analyser certains projets nationaux ou internationaux. Ainsi en France le projet « technolangue »[1], lancé par le ministère français de la recherche et des nouvelles technologies, indiquait parmi ses quatre thèmes d'appel à proposition un volet sur les ressources linguistiques dont l'objectif était de *stimuler la production, la validation et la diffusion de ressources linguistiques pour répondre aux besoins minimaux pour l'étude de la langue française, favoriser la réutilisabilité de ces ressources et diminuer le coût du « ticket d 'entrée » dans le secteur*[2]. Le nombre de projets soumis sur ce volet, associant chercheurs et industriels, montre l'importance de ce thème. Les besoins sont en effet très diversifiés : que ce soit en terme de types de textes (littéraires, scientifiques ou techniques, mono et multilingues), ou en termes d'usages (industriels,

[1] http://www.recherche.gouv.fr/appel/2002/technolangue.htm.

[2] Le coût de développement de ressources textuelles est important et demeure souvent un frein pour de nombreuses études sur notre langue.

professionnels ou grand public), la nécessité de vastes corpus normalisés, annotés et validés s'impose.

2.0 Dictionnaires et lexiques

Le second type de ressources concerne les dictionnaires, les lexiques et les bases de données terminologiques. Bon nombre des arguments développés ci-dessus peuvent aussi s'appliquer en ce domaine. Or aucune exploitation de corpus et aucun traitement automatique de connaissances ne peut se passer du niveau lexical, et la disponibilité de ressources de ce type est unanimement reconnue comme indispensable. Là encore les besoins sont très divers : dictionnaires spécialisés et dictionnaires généraux de langue, lexiques techniques ou bases terminologiques, par exemple, dans un contexte mono ou multilingue.

Si, une fois de plus, la toile offre des réponses diversifiées à ce besoin, nombre de questions demeurent concernant tout à la fois la qualité, la richesse, la couverture et la disponibilité de telles ressources. Il suffit pour s'en convaincre d'analyser les réponses que l'on peut obtenir après une interrogation de la toile à partir, par exemple, de « dictionnaire + langue française » !

3.0 Outils de gestion de ressources textuelles

Un troisième type de ressources, complément des deux précédents, concerne les outils d'exploitation de telles ressources. Deux types d'outils méritent une attention toute particulière :

1. Les outils de gestion et d'accès aux ressources textuelles, lexicales ou dictionnairiques. Que seraient en effet des ressources textuelles ou dictionnairiques du type de celles envisagées ci-dessus sans les logiciels d'exploration de ces ressources ?

2. Les outils de base de traitement de la langue, indispensables pour permettre des recherches plus centrées sur le contenu que la forme, ils doivent être disponibles pour notre langue pour éviter de sans cesse réinventer la roue dans des domaines tels que la lemmatisation, la conjugaison ou l'étiquetage morpho-syntaxique.

Une fois de plus on ne peut que noter, tout en le regrettant, le manque de disponibilité d'outils fiables et généraux de ce type. Faute de cette disponibilité, la première tâche d'une équipe de recherche ou de développement travaillant sur l'exploitation de corpus consiste souvent aujourd'hui à redévelopper de tels outils !

4.0 Intérêt de ressources de références

En conclusion de ce paragraphe, je voudrais faire partager ma conviction de la nécessité de disposer, pour chaque communauté linguistique, de ressources de références (corpus textuels, dictionnaires et lexiques, outils d'exploitation de ces ressources) pour la construction de

modèles ou modules de traitement, pour leur validation et pour leur comparaison. Sans vouloir aller jusqu'à une taylorisation[3] complète de la recherche, il convient en effet de noter qu'aucune équipe de recherche n'est à même de réunir toutes les compétences pour définir et construire de telles ressources que l'ingénierie des langues (Pierrel 2000) est aujourd'hui capable de proposer. La mutualisation de ces ressources et outils de base s'impose donc.

3 L'ATILF propose un ensemble de ressources de référence pour la langue française

Notre laboratoire ATILF, spécialisé dans le domaine de l'analyse et du traitement informatique de la langue française, offre un ensemble de ressources informatisées (Bernard et coll. 2001) composé de bases textuelles et dictionnairiques qui tentent de répondre à ce besoin impérieux de ressources de référence à travers en particulier le Trésor de la Langue Française informatisé (TLFi), la base textuelles Frantext et la plate-forme Stella d'exploitation de ces ressources.

1.0 Le Trésor de la Langue Française informatisé : www.atilf.fr/tlfi

1.0.0 Présentation générale

Le TLFi (Dendien et Pierrel 2003) s'appuie sur Le Trésor de la Langue Française (TLF), dictionnaire de langue réalisé entre le début des années 60 et le milieu des années 90 (CNRS 1976-1994), par l'Institut National de la Langue Française, prédécesseur à Nancy de notre laboratoire ATILF. Dans son récent ouvrage sur les dictionnaires de la langue française, Jean Pruvost 2002 présente ainsi cet ouvrage : *« Ce projet, qui correspond à une entreprise publique ayant requis une centaine de chercheurs pendant 30 ans, avec un dépouillement de plus de 3 000 textes littéraires, scientifiques et techniques, a bénéficié des compétences nationales et internationales les plus éminentes […] Il en résulte, au-delà de la très grande qualité scientifique des articles, une description du fonctionnement de la langue qui ne manque pas d'être impressionnante : 23 000 pages, 100 000 mots, 450 000 entrées, 500 000 citations précisément identifiées. Le TLF relève pleinement d'une lexicographie philologique et historique, recourant aux citations-attestations qui permettent de fonder toutes les analyses morphologiques et sémantiques ».*

Le TLFi se présente à la fois comme une base lexicologique et une base de connaissances dont l'accessibilité est immédiate sous support CDROM (ATILF 2004) ou via Internet (www.atilf.fr/tlfi). Il se distingue des autres dictionnaires électroniques 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 :

[3] *L'organisation taylorisée ne comporte pas nécessairement la pratique de règles, de lois, d'inventions d'une subtilité ou d'une transcendance particulières ; elle consiste à réaliser une combinaison rationnelle des éléments capables d'influer sur le rendement* (PETHOUD, *Organ. industr. et comm.*, 1931, p. 48).

- consultation simple du dictionnaire, article par article, avec mise en évidence ou non de tel ou tel type d'information (définition, exemple…) ;

- consultation transversale, par une requête élémentaire utilisant certains critères (indicateur d'emploi, domaine d'usage…) ;

- consultation plus complexe croisant plusieurs critères. Ces requêtes peuvent être élémentaires ou multi-objets (on peut par exemple extraire tous les syntagmes construits avec le mot *bien* et ayant une définition particulière en droit ; résultats obtenus : *administrateur de biens, aliénation de biens, bien alliéné, bien allié, biens allodiaux, bail d'un bien, cession de biens, biens commun, biens communaux, biens corporels*… en tout plus d'une centaine de réponses).

2.0.0 Spécificités du contenu

Le TLFi se distingue aussi de la plupart des autres dictionnaires informatisés par la richesse de son matériau et la complexité de sa structure :

- originalité de sa nomenclature (incluant syntagmes définis, préfixes, suffixes et autres éléments formants) : c'est en tout 100 000 mots avec leur étymologie et leur histoire, et 270 000 définitions ;

- richesse des objets métatextuels (vedettes, codes grammaticaux, indicateurs sémantiques ou stylistiques, indicateurs de domaines, définitions, exemples référencés…) ;

- richesse des 430 000 exemples, tirés de deux siècles de production littéraire française ;

- diversité des rubriques : une rubrique *synchronie* couvrant la période 1789 à nos jours, une rubrique *étymologie* et *histoire*, et une rubrique *bibliographie* pour les principaux les articles.

3.0.0 Spécificités du 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 a ainsi été introduite 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 "droit" apparaissant au niveau B, par exemple, signifie clairement que les éventuelles subdivisions hiérarchiques du B traitent du droit. Et une définition trouvée dans le paragraphe b) appartenant au paragraphe 2) qui appartient au B introduit nécessairement un sens usité dans le domaine du droit. De manière générale, il est donc possible d'introduire systématiquement une relation entre deux objets X et Y : X étant hiérarchiquement inférieur, égal ou supérieur à Y.

Il est 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. A titre indicatif une telle requête fournit en résultat 49 verbes ayant trait au maniement des voiles en marine.

La richesse du balisage du TLFi (structuration et hiérarchisation d'une quarantaine d'objets méta textuels différents avec jusqu'à 23 niveaux hiérarchiques grâce à l'introduction dans le texte de plus de 36 millions de balises XML) permet l'identification de nombreux types d'objets et leur mise en relation hiérarchique, et contribue à obtenir des résultats d'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.

2.0 La base Frantext de données textuelles : www.atilf.fr/frantext

1.0.0 Présentation générale

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

Historiquement (Bernet et Pierrel 2005), le but premier du corpus textuel 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 "bien", par exemple, se trouvait ainsi doté d'une concordance systématique de ce mot, triée suivant différents critères. 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 estimer que ce corpus textuel a fourni environ 90% des quelques 430 000 exemples cités dans le TLF.

Dans les années soixante-dix, la constitution des dossiers de mots, ainsi que les extractions des exemples finalement retenus, étaient assurées par de lourds logiciels non interactifs 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 un 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.

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 et doté d'un outil d'interrogation de plus en plus efficace.

2.0.0 Etat actuel du corpus

Le corpus actuel comporte 3747 textes (plus de deux milliards de caractères) dont les dates s'échelonnent de 1507 à 1998. Le corpus textuel de Frantext vient prolonger un corpus de français médiéval et de moyen français (environ 300 textes couvrant les années 847 à 1502), offrant ainsi l'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. A ce jour 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 qui se trouvait encore dans l'état de la saisie initiale (début des années 60) et dont certains textes avaient été victimes d'avatars liés à la technologie de l'époque ont fait l'objet d'une campagne intensive de correction au cours des cinq dernières années.

- Qualité des éditions saisies : l'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).

- Complémentation du corpus : notre politique veut 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, chaque année, de la totalité des textes inscrits au programme de l'agrégation.

Frantext est disponible sur internet, par abonnement, sous deux formes :

1. la totalité du corpus (3747 textes) que l'on peut interroger sur les formes graphiques du texte.

2. 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 peut être interrogée à la fois sur les formes graphiques et sur les catégories grammaticales.

Cette base est complétée par corpus technique constitué de 1083 normes AFNOR, mis à la disposition des chercheurs, en libre accès via une interface identique à Frantext (www.atilf.fr/tilt). Cette base complémentaire correspond à l'un des résultats du projet TILT (Trésor informatisé de la Langue Technique), faisant partie du projet AGILE dans le cadre de l'action Technolangue, action commune aux trois réseaux de recherche et d'innovation technologique (RNRT, RNTL, RIAM), financée dans un cadre interministériel français.

3.0 Une intégration de l'ensemble grâce à un même outil d'exploitation : le logiciel Stella

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 ressources complémentaires telles les versions informatisées du dictionnaire de l'Académie française, gérées elles aussi par Stella : 8e édition de 1935 www.atilf.fr/academie8 et 9e édition (en cours) www.atilf.fr/academie9 .

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

- Divers utilitaires incluant des tris, le traitement des expressions régulières, et surtout une base de données fondée sur la nomenclature du TLF permettant des opérations de flexion ou de lemmatisation.

- Interface Web permettant la mise en œuvre facile d'interfaces utilisateur, des fonctions de gestion de "sessions utilisateur " 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 la forme *s'assujettit* permet d'afficher l'article *assujettir* du TLFi ou des 8e ou 9e édition de l'Académie.

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

La conception de STELLA, en tant que logiciel de gestion de base textuelle, repose sur des principes mathématiques rigoureux (Dendien, 1991) et intègre un nombre important de fonctionnalités.

1.0.0 Fonctions de " bas niveau "

Ces fonctions permettent de construire 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 prévisibles par la Théorie de l'Information, avec cette 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 (en ne dépendant 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 du système d'indexation assure des performances optimales en minimisant les échanges entre mémoire de masse et mémoire centrale.

2.0.0 Fonction de " haut niveau "

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

- Tout objet textuel est manipulable par des méthodes-standard (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 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 entrer 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 en 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 permettant 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 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 permettent d'introduire facilement un " savoir linguistique ". Par exemple, rien n'est plus facile que d'introduire, si on le désire, les déclinaisons latines.

4 Exemples d'exploitation de nos ressources sous Stella

1.0 Recherche dans le TLFi

Bien que le logiciel Stella soit doté d'un compilateur de langage de requête très avancé, le TLFi présente un certain nombre de spécificités qui nous ont conduits à définir un langage de requête spécifique au TLFi et à réaliser son compilateur. Le langage de requête définit un vocabulaire associant un mot-clé à chaque élément XML. Par exemple, le mot-clé définition est associé à l'élément DEF de la DTD. Cette association est réalisée à l'aide d'un simple fichier de correspondance, très facile à mettre à jour avec un simple éditeur de texte. Il est ainsi très aisé, par exemple, de réaliser un jeu de mnémoniques adaptés à des utilisateurs anglophones ou hispanophones, et aussi de permettre ou d'interdire la visibilité de tel ou tel élément XML.

1.0.0 Principes du langage de requête

Le principe du langage de requête du TLF est simple. La requête :

```
" X:domaine(agriculture);Y:définition(instrument);
Y i (X);Z:source(Académie);Z i Y; "
```

s'interprète de la manière suivante :

- *soit X un indicateur de domaine technique contenant le mot agriculture,*
- *soit Y une définition contenant le mot instrument,*
- *Y est inclus dans la portée de X (on remarquera la notation " Y i (X) " qui signifie que l'élément X est inclus dans la portée de l'élément Y, alors que " Y i X " signifie que l'élément X est inclus dans l'élément Y). Cette clause (cf. la notion de portée des objets) implique que la définition est valable dans le domaine de l'agriculture,*
- *soit Z une source (bibliographique) contenant le mot Académie,*
- *Z est inclus dans Y (cette clause implique que la définition est empruntée au dictionnaire de l'Académie française).*

Cette requête va déclencher la recherche de tous les triplets (X, Y Z) respectant le système de contraintes énoncé. Elle peut se paraphraser ainsi : *" Chercher, dans le domaine de l'agriculture, les définitions relatives à un instrument et empruntées au dictionnaire de l'Académie "*.

On remarquera l'aspect non procédural du langage de requête qui permet de décrire un système de contraintes à résoudre, mais pas les opérations nécessaires pour y parvenir. On voit, dans cet exemple que le langage de requête met en jeu à la fois le type des objets textuels, leur contenu textuel éventuel, et les relations entre les objets de la requête. La différence entre les deux types de relation (inclusion et dépendance hiérarchique) est fondamentale et constitue le seul point délicat à appréhender pour une bonne manipulation du langage de requête.

Pour conclure cette présentation rapide du langage de requête du TLFi, il convient de comprendre la notion de connexité d'une requête : toute requête introduit un ensemble d'éléments ({X, Y, Z} dans notre exemple) muni d'une relation binaire, et peut donc être considéré comme un graphe G. Si le graphe G est non connexe, cela signifie que la requête demande de rechercher au moins deux sous-ensembles d'éléments sans aucun lien logique l'un avec l'autre. Une telle requête n'a évidemment aucun sens, et sera rejetée par le compilateur.

2.0.0 Les différentes possibilités d'exploitation du langage de requête

Comme nous venons de le voir, une requête permet de décrire ce qu'il faut chercher. Il reste encore à exprimer la manière dont les différents résultats trouvés doivent être restitués à l'auteur de la requête. Ce dernier point dépend du contexte dans lequel la requête a été élaborée : exploitation à distance via un serveur de requêtes ou exploitation sous contrôle d'une interface pour le Web.

1.0.0.0 Exploitation à distance via un serveur de requêtes

Ce mode consiste à utiliser Internet pour poster une requête à une application résidant sur le serveur gérant le TLFi. La requête est empaquetée dans une coquille XML. Il est nécessaire de compléter la requête proprement dite avec des clauses exprimant quelles informations on veut obtenir.

Exemple de requête à poster au serveur :

```
<TLFquery>
     <query>
          X:domaine(agriculture);Y:définition(instrument);
          X i(Y);Z:source(Académie);Z i Y;
     </query>
     <extract>
          Y
     </extract>
</TLFquery>
```

Réponse du serveur :

```
<TLFanswer>
     <sol n="1">
```

```
            <Y>
            texte de la première définition
            </Y>
        </sol>
        <sol n="2">
            <Y>
            texte de la seconde définition
            </Y>
        </sol>
        --------etc.---------
</TLFanswer>
```

2.0.0.0 Exploitation sous contrôle d'une interface pour le Web

Dans ce mode, l'utilisateur ne manipule pas directement le langage de requête. En effet, l'interface graphique lui propose différents formulaires de recherche. Lors de la soumission du formulaire, les données de l'utilisateur sont collectées et automatiquement transformées en une expression de requête. On trouvera dans (Bernard et coll. 2002) une présentation explicitant les divers usages possibles de cette interface web.

L'interface offre des possibilités graduées :

- recherche simple : elle consiste à rechercher les articles concernant le mot introduit par l'utilisateur ;

- recherche assistée : elle propose un formulaire de recherche permettant à l'utilisateur d'imposer des contenus textuels à des types d'objets donnés, le logiciel prenant en charge de manière transparente les relations hiérarchiques entre objets ;

- recherche complexe : elle propose un formulaire permettant d'exprimer une requête avec une puissance comparable à la manipulation directe du langage de requête. L'utilisateur peut explicitement spécifier les types des objets recherchés, leurs contenus textuels éventuels et leurs relations hiérarchiques.

Dans tous ces cas, lors de la soumission du formulaire, le logiciel collecte les informations, les transforme en une requête qui est soumise au compilateur, puis exécutée.

Les résultats de la recherche peuvent être affichés suivant deux modes différents :

- mode global : supposons qu'une recherche mette en jeu N objets textuels. Chaque résultat est donc constitué d'un N-uplet d'objets. Le mode global consiste à afficher tous les résultats. Pour chaque résultat est affiché le contenu textuel de chaque objet du N-uplet, ou éventuellement, au choix de l'utilisateur, le sous-ensemble du N-uplet constitué par les objets qui l'intéressent le plus ;

- mode “en contexte” : il consiste à afficher les résultats un par un. Le texte de l'article dans lequel le N-uplet a été trouvé s'affiche intégralement. Le début et la fin du texte correspondant à chaque objet du N-uplet sont matérialisés par de petites images colorées et numérotées de 1 à N. Dans le cas où toute la totalité de l'article n'est pas visible à l'écran (ceci arrive fréquemment en raison de la taille importante des articles du TLF), l'utilisateur dispose de boutons de navigation lui permettant de

centrer l'affichage immédiatement sur l'objet du N-uplet qui l'intéresse le plus particulièrement.

2.0 Les recherches complexes dans Frantext

1.0.0 Principes de base du travail sur Frantext

On peut distinguer deux phases fondamentales dans le travail : choix des textes que l'on veut étudier, puis série d'études portant sur les textes choisis. La recherche débute donc par la *création du corpus de travail*. Il est ensuite possible de le *visualiser*, d'*effectuer des recherches* puis de *télécharger les données.*

1ère étape : La sélection du corpus de travail. Elle peut s'effectuer sur l'ensemble des textes de la base, un ou plusieurs auteurs, une œuvre ou un ensemble d'œuvres, un ou plusieurs genres littéraires, une tranche chronologique ou la combinaison de plusieurs de ces critères.

2ème étape : les requêtes sur le corpus de travail. Certaines sont simples à mettre en œuvre, d'autres, plus compliquées. Une aide en ligne accompagne chaque service.

Recherches simples d'une graphie, d'une phrase ou de plusieurs graphies en cooccurrence dans une ou plusieurs phrases.

Recherches plus complexes, dont voici une liste non exhaustive :

- les formes fléchies d'un verbe, d'un substantif ou d'un adjectif ;
- des formes tronquées : mots suffixés en " *ette* " par exemple ;
- des listes de mots ;
- des expressions à choix multiples en une seule requête. Par exemple "*maison (blanche | verte | jaune)*" recherchera *maison blanche, maison verte ou maison jaune* ;
- des statistiques sur la fréquence des mots ;
- des études de vocabulaire autour d'un mot pivot (par ex. le vocabulaire employé dans les phrases contenant le mot "*droit*") ;
- des phénomènes linguistiques complexes tels que la quantification, les constructions pronominales, les temps composés, etc. ; ce type de recherche est rendu possible grâce à l'écriture de grammaires formelles constituées de règles paramétrables. Ces grammaires permettent de rechercher des contextes arbitrairement complexes sur le corpus.

Il est possible au terme de sa recherche de télécharger les résultats. Les résultats sont encodés suivant la norme ISO 8859-1, compatible avec les usages en vigueur sur les principaux systèmes (MS-Windows, Unix, Mac-OS, …). La longueur des contextes obtenus dans la base Frantext est limitée à trois pages pour les textes libres de droits, et à une courte citation de 300 signes pour les textes protégés par des droits d'auteur ou d'éditeur.

2.0.0 Les principes de recherche

Les critères de recherches peuvent être saisis au cours d'une session (recherche simple avec saisie d'*expressions de séquence*) ou saisis dans une liste préparée avant d'effectuer une

recherche grâce aux outils "*création/édition de listes de mots*", ou aux "*grammaires*" prédéfinies.

1.0.0.0 Les règles de saisie d'une expression de séquence

Une expression de séquence a pour but de rechercher des contextes contenant une suite de mots consécutifs (d'où le terme séquence). La saisie d'une expression de séquence est effectuée sur le formulaire disponible sous l'option "*recherche dans les textes*". La saisie d'une séquence consiste à saisir une suite de mots consécutifs. Une expression de séquence est constituée de N sous-expressions :

- opérateur ET, matérialisé par la juxtaposition de deux mots : ex. *maison blanche* ;
- opérateur OU, matérialisé par la barre verticale | et les parenthèses : (*maison*|*palais*) *d'un (blanc*|*bleu)sale* va chercher *maison d'un blanc sale, maison d'un bleu sale, palais d'un blanc sale, palais d'un bleu sale* ;
- verbe fléchi &c : *&cassujettir* ira chercher toutes les formes du verbe *assujettir*, conjuguées ou non ;
- substantif ou adjectif fléchi &m : *&mbail* va chercher *bail, baux*, *&mgrand* va chercher *grand, grande, grandes, grands* ;
- sous-expressions optionnelles & ? : *un & ?grand homme* cherchera *un homme* ou *un grand homme* et *un & ?(très grand) homme*, *un homme ou un très grand homme* ;
- Mot quelconque &q introduisant une discontinuité dans la séquence *: Un &q homme* cherchera une séquence *Un (…) homme* avec un mot entre *un* et *homme*, *Un &q(0,2) homme* cherchera une séquence avec 0, 1 ou 2 mots entre *un* et *homme* ;
- La négation ^ : *homme ^très grand* qui est une expression de séquence qui va chercher tous les contextes du genre *homme XXX grand* qui tels que XXX ne soit pas égal à *très*. ; *&carriver ^&e(g=A)* à est une expression de séquence qui va sélectionner des contextes tels que *arriveront vite à*, mais pas un contexte tel que *arrivaient en même temps à*. En effet *^&e(g=A)* désigne une graphie qui n'est pas un adjectif.

2.0.0.0 Exemples d'utilisation de grammaires dans Stella

La possibilité, offerte dans Stella, 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 un exemple, l'étude des utilisations pronominales 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 et entre crochets, ex : [*commentaire*], et 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 l'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 de la rendre paramétrable pour fonctionner avec un verbe quelconque.

3.0 Autres 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 suivant :

1.0.0 Recherche de cooccurrences et collocations:

La taille du corpus 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 ou les locutions si on est intéressé par un travail sur la phraséologie.

2.0.0 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 domaine 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 (le droit, la marine, mais aussi la mythologie ou l'œnologie) pour peu que les indicateurs correspondants figurent dans les articles du TLF papier.

3.0.0 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.

4.0.0 Etudes de syntaxe locale

Pour repérer des motifs récurrents, des différences syntaxiques ou d'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 (plus de 1200 solutions) pour étudier leur environnement à partir des exemples correspondant 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.

5.0.0 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 plus de 240 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 plus de 50 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.

6.0.0 Etudes de stylistique

Le TLFi permet également de demander tous les exemples de Colette contenant une forme conjuguée du verbe aimer dans le corps de l'exemple. Et Frantext est un outil performant pour les études des vocabulaires spécifiques à un auteur. Le TLF a pris en compte les néologismes

et les 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.

7.0.0 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 les domaines médical philosophique, technique ou autre. Les étymologistes peuvent donc être intéressés par sa consultation, de même que des phonéticiens et phonologues par les rubriques Prononciation. Certains articles, particulièrement ceux concernant les mots grammaticaux, sont riches de remarques concernant les indications d'emploi..

5 Conclusion

Frantext, TLFi et le logiciel Stella forment un ensemble de ressources de référence, 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 coll. 2001).

Ces ressources, mises à la disposition de la communauté scientifique, peuvent être le fondement de projets divers, 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. Conscient 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 scientifique à travers des projets coopératifs de recherche.

Remerciements

Au terme de cet article, je tiens à exprimer mes plus vifs remerciements à tous les collègues de l'INaLF et de l'ATILF d'aujourd'hui sans qui ces ressources n'existeraient pas. Merci donc à l'ensemble des rédacteurs du TLF et à celles et ceux qui ont élaboré les versions informatiques de ces ressources, en particulier Jacques Dendien pour le logiciel Stella, l'équipe de réalisation du TLFi et le service des Bases Textuelles du laboratoire.

Références

ATILF (2004) *Trésor de la Langue Française informatisée*, CNRS Editions, 2004 (Livre 592 p. et CDROM)

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.

Bernard P., Dendien J., Lecomte J. et Pierrel J.M. (2002) « Un ensemble de ressources informatisées et intégrées pour l'étude du français : FRANTEXT, TLFi, dictionnaires de l'Académie et logiciel Stella, présentation et apprentissage de leurs exploitations », *Actes de TALN 2002*, vol 2, p. 3-36, Nancy, 24-27 juin 2002 , disponible aussi à l'adresse : http://www.loria.fr/projets/TALN/TALN/index.html.

Bernet C. et Pierrel J.M. (2005) « Histoire de Frantext : constitution d'une base textuelle (1964-2002) et perspectives », in *L'édition électronique en littérature et dictionnairique : évaluation et bilan*, Presses Universitaires de Rouen, Editions Champion, à paraître.

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*, Didier Erudition, Paris, pages 25-34.

Dendien J. et Pierrel J.M. 2003 « Le Trésor de la Langue Française Informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence », *Traitement Automatique des Langues*, Vol 44, Editions Hermès.

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, Armand 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.

Pruvost J. 2002 *Les dictionnaires de la langue française*, collection Que Sais-je ?, PUF, Paris.

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