S'approprier des instruments d'observation de la langue pour élaborer des recherches : le TLFi et Frantext pour des étudiants de linguistique

Cécile Fabre, Michelle Lecolle

S'approprier des instruments d'observation de la langue pour élaborer des recherches : *le TLFi* et *Frantext* pour des étudiants de linguistique

Cécile Fabre, CLLE-ERSS, Université de Toulouse le Mirail

Michelle Lecolle, CELTED, Université Paul Verlaine-Metz

Introduction

Dans cet article, nous proposons une réflexion née de plusieurs expériences d'enseignement[1], au cours desquelles nous avons cherché à intégrer dans le cursus d'étudiants en linguistique la maîtrise et l'utilisation d'outils informatiques permettant l'exploration de ressources linguistiques – corpus (*Frantext*) ou dictionnaire informatisé (*Trésor de la Langue Française informatisé*). Ces enseignements ont pour objectif, au-delà de la familiarisation avec les objets « dictionnaire » et « corpus », l'acquisition par les étudiants d'un savoir-faire leur permettant de constituer des données linguistiques destinées à alimenter leurs travaux personnels, et de développer leurs capacités de description et d'analyse de ces données.

À travers cet exposé, nous chercherons à montrer la part que peut prendre, pour un étudiant, une expérience de recherche empirique sur des données dans un travail d'observation et de conceptualisation à partir de la langue et des discours. Selon la formule de Rastier (2005), « les données sont ce qu'on se donne » : ainsi, c'est bien le travail de constitution, de classement et de formalisation des observables à partir de faits linguistiques issus de textes ou d'énoncés lexicographiques qui est considéré ici comme le point de départ de la recherche. Pour isoler et systématiser ce qui sera ensuite considéré comme pertinent, l'étudiant doit en effet acquérir et aiguiser une pratique réflexive et introspective à partir de sa propre langue. Dans cette perspective, apprentissage technique d'un outil automatisé et observation des données langagières vont de pair et s'alimentent mutuellement. Les activités d'exploration du dictionnaire informatisé et de corpus amorcent ainsi le processus de description que décrit Martin, fondé sur la collecte de données et la sélection d'observables :

> « La première tâche du linguiste est d'observer et de décrire : son étude portant sur un objet du monde qui préexiste à son investigation, celui-ci se prête, par nature, à un traitement empirique. A vrai dire, cette question en présuppose une autre qui est : « Quoi décrire ? » ; « Quels sont les faits qui se prêtent à la description et comment les collecter ? » (Martin 2002 :19)

La confrontation à des données langagières qu'il a lui-même collectées amène l'étudiant à s'interroger sur les caractéristiques de l'objet linguistique qu'il cherche à circonscrire, et, pour reprendre à nouveau les termes de Martin, à « applique[r] aux faits collectés des principes […] de description » pour esquisser une « synthèse descriptive ».

Nous présentons successivement les activités que nous avons conçues autour du *Trésor de la Langue Française Informatisé* (désormais *TLFi*) et de *Frantext* pour initier cette démarche d'analyse. Nous commençons dans les deux cas par exposer les principales fonctionnalités des deux environnements de travail, les compétences générales qui sont sollicitées par l'appropriation de ces instruments, avant de montrer à partir d'exemples de recherche précis comment s'élabore le travail de description linguistique.

---

[1] Aux départements de sciences du langage de l'Université de Toulouse-Le Mirail et de l'Université Paul Verlaine de Metz. Certains des exemples présentés sont tirés de travaux d'étudiants de L3 de Sciences du Langage des promotions 2006-2007, 2007-2008 et 2008-2009 de l'UPVM. Nous les remercions pour leur collaboration (involontaire) au présent article.

1. Le *TLFi*[2] : découverte de nouveaux modes d'exploration des données lexicales

Les dictionnaires informatisés sont des supports utiles pour qui veut susciter l'intérêt des étudiants vis-à-vis des dictionnaires, et initier de nouvelles démarches pour exploiter les ressources qu'ils contiennent. Le support informatique change le regard porté sur les dictionnaires, et facilite la découverte. Au-delà du seul *TLFi*, l'accès à de nombreux dictionnaires numérisés via le site du laboratoire ATILF ou celui du projet ARTFL[3] qui met en ligne des dictionnaires de différentes époques, rend possibles des travaux pratiques sur machine et crée les conditions d'un travail d'observation différent. Une séance de travaux dirigés combinant la découverte de ces deux sites permet de mesurer avec précision et dans toutes ses dimensions la « révolution électronique » dont parle Jean Pruvost (2006 :153). En effet, le balisage de différents champs de description (les objets textuels) multiplie des modes d'accès au contenu ; l'expression de la requête est facilitée par un ensemble de fonctionnalités (troncature, approximation orthographique, calcul phonétique) ; des cheminements nouveaux sont suscités par des liens hypertextuels internes (parcours analogiques) ou externes (accès à des ressources complémentaires – *Frantext* et dictionnaires de l'Académie notamment). Enfin, les modes de visualisation se diversifient (listage des entrées, concordances, mise en relief des zones pertinentes par des codes de couleurs…). Ces environnements créent les conditions d'un renouveau de l'intérêt pour les dictionnaires et offrent des pistes nouvelles pour leur exploitation. Les étudiants sont alors prêts à un apprentissage technique relativement rébarbatif parce qu'ils peuvent aussitôt le mettre en pratique et en saisir les bénéfices pour leur formation en linguistique. Le *TLFi* offre sans conteste les potentialités les plus riches parmi les dictionnaires du français pour engager cette démarche d'exploration et de collecte de données lexicales.

1.1. Utilisation avancée du TLFi

Le *TLFi* donne accès à l'ensemble des informations que renferme le *TLF* lui-même, mais sous le jour particulier que permet une recherche automatisée. Nous en citons pour mémoire quelques particularités, l'essentiel des possibilités offertes par le *TLFi* étant détaillée dans Dendien et Pierrel (2003).

Tout d'abord, à l'ordre alphabétique du dictionnaire est substituée une interrogation directe par mot-vedette, ou par le biais d'autres champs d'information rendus accessibles grâce à une structuration et un balisage systématiques des articles. Ceux-ci sont en effet décomposés en « objets textuels » (jusqu'à 40) correspondant aux éléments traditionnels de composition d'une microstructure : mot-vedette, catégorie et morphologie, définition, exemples[4], indicateurs de domaine, indicateurs grammaticaux, sémantiques et « stylistiques », synonymes et antonymes, constructions, etc. C'est cette structuration des 100 000 entrées (vedettes ou sous-vedettes) du *TLFi* qui permet alors à l'utilisateur d'accéder directement à des informations ciblées sur les contenus de certains objets, de systématiser ses requêtes et de les croiser. La recherche peut ainsi concerner plusieurs objets textuels entretenant entre eux des liens hiérarchiques de *dépendance* (lorsqu'un objet est situé dans la portée d'un autre objet) ou d'*inclusion* (dans le cas d'objets composites comme l'entrée ou l'exemple). A titre d'exemple, le « contenu » *musique (mus.)* dans l'objet textuel DOMAINE TECHNIQUE peut être croisé avec le « contenu » *substantif* dans l'objet textuel CODE GRAMMATICAL pour accéder à l'ensemble des substantifs dont un des sens au moins relève du domaine de la musique. De

[2] Une partie des éléments présentés dans cette section a fait l'objet d'une communication dans le cadre du colloque « Lexicographie et informatique : bilan et perspectives », organisé par l'ATILF à l'occasion du 50ème anniversaire du lancement du TLF, en 2008.

[3] http://www.atilf.fr/atilf/res_ling_info.htm et http://www.lib.uchicago.edu/efts/ARTFL/projects/dicos/

[4] L'exemple est lui-même structuré en plusieurs objets textuels comprenant le texte lui-même, les sources et les auteurs des sources.

fait, la structuration de la ressource et l'outil de recherche introduisent la possibilité d'accéder très facilement à une information complexe perdue au cœur des dictionnaires papier.

La formulation de contenus textuels spécifiques permet par ailleurs d'exprimer des contraintes sur le positionnement des éléments ciblés dans l'objet (place dans la définition, par exemple – voir *infra* pour des illustrations), d'obtenir toutes les formes fléchies d'un même lexème, d'exclure certains types d'éléments, d'intégrer des éléments non connus *a priori* (mots quelconques), de gérer des listes constituées manuellement par l'usager ou extraites automatiquement du dictionnaire sur la base de critères graphiques. On peut ainsi par exemple obtenir la liste des substantifs se terminant par *-age*, et restreindre les résultats à ceux dont la définition débute par *action*. L'utilisateur disposera alors facilement d'un corpus de mots sur lequel exercer sa sagacité.
L'utilisation avancée du *TLFi* sur laquelle nous basons nos pratiques d'enseignement consiste à utiliser pleinement les fonctionnalités du mode de « recherche complexe »[5] (Dendien et Pierrel 2003 : 27). Pour des usagers novices, cela passe par un apprentissage technique assez exigeant, qui suppose avant toute chose d'acquérir une bonne connaissance des objets textuels manipulables. Il est néanmoins rapidement payant. A l'issue d'un apprentissage de quelques heures, les étudiants sont à même de mener de petites études linguistiques portant sur des données qu'ils auront eux-mêmes construites.

Un enseignement articulé autour de la découverte du *TLFi* permet ainsi de développer diverses compétences complémentaires et facilite plusieurs types d'initiation :

Des compétences techniques : il s'agit d'apprendre à formuler les requêtes, ce qui passe par l'apprentissage d'une syntaxe particulière dans la formulation des « contenus », et surtout par la maîtrise de l'expression des liens logiques entre objets textuels (Pierrel 2003). Au-delà, le *TLFi* est un point de départ profitable pour amener les étudiants à la compréhension de certaines procédures informatiques qui constituent la base des traitements appliqués aux données langagières : lemmatisation (permettant d'obtenir toutes formes fléchies d'un même lexème), création de listes de mots à partir de critères proches des expressions régulières, sensibilisation aux techniques de balisage de textes structurés, etc. ;

Des compétences lexicographiques : la maîtrise des aspects techniques amène les étudiants à acquérir une connaissance approfondie de la microstructure du TLF. Il s'agit de connaître les différents objets qui la structurent (définition, indicateur, construction syntaxique, crochets...), les liens hiérarchiques qu'ils entretiennent, et de savoir par quels types d'objets textuels une information donnée est codée. Un contact préalable avec la version papier du dictionnaire s'avère utile ;

Des compétences linguistiques : l'objectif premier de cet enseignement est d'amener l'étudiant à avoir plus tard le réflexe de puiser dans le *TLFi* certaines des données dont il a besoin. Cette démarche s'appuie sur les enseignements donnés dans des modules de linguistique générale, en particulier de lexicologie, et les renforce. Les étudiants peuvent notamment obtenir des dérivés d'un lexème base, étudier la valeur sémantique des suffixes (en observant par exemple les indicateurs associés aux adjectifs en *-eux*, les domaines couverts par les noms en *-ose*), les différentes dimensions de variation du vocabulaire (vocabulaire argotique, familier, usages régionaux...), l'organisation sémantique du vocabulaire (en rassemblant par exemple le vocabulaire qui concerne la prison, ou qui a trait à l'activité de manger, etc.) ;

Des compétences méthodologiques : le recours à l'outil informatique, désormais banalisé chez les étudiants, est mis en perspective et questionné de diverses manières par ce type d'utilisation. Ils doivent apprendre à ne pas tout attendre de l'outil, à combiner requêtes

[5] Pour des exemples de recherche complexe, voir Tableau 1 et Tableau 2 *infra*.

automatiques et filtrage manuel pour se conformer à un réel objectif de recherche. Une réflexion concrète sur les notions de précision et de rappel des requêtes[6] est très bénéfique. Comment évaluer la performance d'une requête ? Comment s'assurer qu'elle ne passe pas à côté de résultats pertinents ? Avec le *TLFi*, les étudiants découvrent qu'il ne suffit pas qu'une requête donne un résultat pour qu'on puisse s'en satisfaire, mais qu'il faut observer de très près les résultats et réitérer la recherche. C'est un excellent moyen d'aiguiser le regard sur les données linguistiques et d'acquérir du recul par rapport à leur utilisation courante de l'informatique.

1.2. Un instrument pour réunir des matériaux lexicaux

En somme, comme nous le soulignerons aussi plus bas à propos des corpus de textes, l'acquisition par les étudiants des compétences citées à propos du *TLFi* permet de les sensibiliser à des objectifs de recherche concrets. S'il n'est pas possible de mener à bien avec des étudiants peu avancés des recherches à grande échelle, l'utilisation du *TLFi* les met néanmoins sur la piste d'expériences de recherches lexicales réalistes, utiles à ceux qui travaillent sur la langue. Nous développons à présent quelques exemples d'expériences réelles qui permettront d'illustrer les va-et-vient et les tâtonnements fructueux qu'occasionne le couplage apprentissage technique/observation des résultats fondée sur la compétence linguistique et l'observation réflexive.

Premier exemple : constituer une liste de mots selon des critères graphiques

Dans le cadre d'une activité d'expression écrite, un professeur des écoles souhaite constituer un lexique des mots qui riment avec *bulle*. Ce lexique doit permettre d'aider ses élèves de CE1 à écrire de petits textes en puisant dans un vocabulaire proposé[7]. Les étudiants ont pu réaliser cette activité dans le cadre d'un travail dirigé : une liste (nommée *ule*) est extraite automatiquement du dictionnaire selon un critère formel permettant de traduire au mieux cette contrainte sur le plan graphique. L'essentiel du travail consiste à construire la liste *ule*. Pour mener à bien la recherche, c'est-à-dire pour éliminer les réponses non pertinentes d'une part (limiter le bruit) et obtenir le maximum de réponses pertinentes d'autre part (éviter le silence), l'étudiant doit étudier tous les paramètres de la question (et, dans les faits, les découvrir progressivement). La recherche de mots qui riment avec *bulle* nécessite que ces mots se terminent par les sons [yl]. Cette première contrainte doit amener l'étudiant à réfléchir conjointement au code graphique et à la phonie : il devra par exemple éliminer en finale les suites graphiques ne se prononçant pas [yl], soit *-eul(e), -oul(e), -aul(e)*. Il devra également enrichir la liste en envisageant les verbes qui, conjugués à certaines personnes du présent de l'indicatif et du subjonctif, riment avec *bulle* (on trouve une centaine de verbes en *uler* – *affabuler* ou *tintinnabuler* par exemple). Les verbes étant présentés en mot-vedette dans le dictionnaire par leur lemme (infinitif en *-er*), il lui faudra penser au *-r* graphique final. En définitive, c'est par un enrichissement progressif que l'étudiant, formé à la codification proposée par le système, accèdera à la (meilleure) solution, représentée par le code de constitution de listes suivant : /.*[^aeo]ull?e?r?/ – élimination des graphies *a, e* et *o* avant *ul* ; possibilité (mais non obligation) de deux *l* puis de *e* et *r*.

Cette liste peut ensuite être utilisée pour effectuer des recherches ciblées. Le Tableau 1 montre une requête en mode « recherche complexe » qui spécifie que les articles recherchés ont pour MOT-VEDETTE un des mots de la liste *ule* (appelée par le code &lule), et pour CODE GRAMMATICAL *substantif*. On décide dans ce cas particulier de récupérer seulement ceux qui

[6] La précision mesure la proportion de réponses satisfaisantes parmi les réponses obtenues ; le rappel mesure la proportion de bonnes réponses obtenues parmi le total de bonnes réponses potentielles que la base lexicale (le dictionnaire) contient.

[7] Cette activité a été inspirée par une séquence pédagogique réelle.

relèvent du DOMAINE TECHNIQUE *zoologie* de manière à pouvoir organiser l'ensemble en sous-listes (on en trouve 41, dont *antennule*, *tarentule*, *tentacule*).

n° d'objet	type de l'objet	lien	contenu
1	Mot-vedette	Inclus dans l'objet *4*	*&lule*
2	Code grammatical	Inclus dans l'objet *4*	*substantif*
3	Domaine technique	Dépendant de l'objet *4*	*zoologie*
4	entrée		

Tableau 1: les substantifs rimant avec bulle et relevant du domaine de la zoologie

Nous avons employé ci-dessus le mot « solution ». De fait, il n'y a souvent qu'une « moins mauvaise » solution, voire plusieurs solutions possibles, et les recherches suivantes l'illustreront. Cette constatation ne doit pas être perçue nécessairement comme un écueil : conscient des difficultés diverses, dues notamment à la complexité de l'objet dictionnaire et de sa structuration, dues également à la complexité de l'informatisation elle-même[8], l'étudiant (in-)formé sera amené à explorer plusieurs voies, à observer les résultats et à les comparer. Il importe en effet que les étudiants ne se contentent pas d'une réponse quantitative (x résultats répondant à un essai de requête) mais jugent dans les faits si ce qu'ils obtiennent correspond bien à ce qui était attendu et envisagent dans le cas contraire des améliorations appropriées de la requête.

Deuxième exemple : aborder des propriétés sémantiques

Nous l'avons dit, le mode d'interrogation auquel donne accès un dictionnaire informatisé et outillé est renouvelé : en simplifiant, on peut dire que le *TLFi* modifie l'ordre des pôles connu/inconnu d'une recherche. Ainsi, s'il est nécessaire, dans un usage traditionnel de dictionnaire papier de connaître au moins la forme du mot par lequel on y entre, il est possible (bien que peu aisé) avec le *TLFi* d'entrer par le sens (connu ou évalué), pour accéder à un mot ou une séries de mots (inconnu). La recherche des objets textuels adéquats pour ce faire est déjà un premier élément de réflexion pour les étudiants – reposant sur une acquisition des finalités particulières de ces différents objets textuels dans une logique lexicographique. Il s'agit en premier lieu de l'objet DEFINITION, lorsque la définition est hyperonymique (par inclusion) ou méronymique (débutant par exemple par le mot *partie*). Un autre mode d'accès, d'ordre onomasiologique, est fourni par l'objet textuel qui rassemble, dans le *TLFi*, synonymes et antonymes. Enfin, des informations sémantiques comme *par métaphore, par métonymie*, *figuré*, recherchées directement dans l'objet textuel INDICATEUR donnent accès, à l'intérieur d'un même article, aux formes de polysémie correspondantes.

C'est en mettant à profit ses connaissances en sémantique lexicale et en lexicographie, et en observant la structure de plusieurs articles que l'étudiant peut apprendre à tirer parti de ces différentes informations pour construire une requête destinée à obtenir au final une liste de mots ayant, par exemple, un même hyperonyme, donc rapprochés par leur classificateur : dans le cas cité des noms d'action en *-age*, la définition de ces noms est supposée débuter par *action*. Dans les faits, tous les noms d'action ne sont pas définis de la sorte ; des requêtes destinées à obtenir les synonymes de l'hyperonyme définissant peuvent alors permettre d'enrichir la liste des incluants. On procède ainsi de manière cumulative : partant à la recherche de noms de qualité par exemple, on a ainsi d'abord recherché les synonymes de *qualité*, pour en faire une liste qu'on a ensuite utilisée dans l'objet DEFINITION. Un regard

[8] Corbin et al. (1995) ainsi que Martin (2001) ont déjà signalé des difficultés plus radicales, notamment de cohérence dans les modes de notations des constructions du verbe, et, à partir de certains noms et adjectifs (adjectifs de couleur), le manque crucial d'indications de classifications sémantiques stables.

critique sur les résultats des différentes étapes est ici crucial : tous les résultats de l'objet SYNONYME/ANTONYME, par exemple, ne sont pas adaptés à une utilisation comme définissant de la série de mots recherchés par ce biais – ainsi, le synonyme *vie active*, donné pour *action*, ne sera pas retenu comme bon candidat à exprimer une définition des 'actions'.

Des étudiants ont utilisé une procédure comparable, mais sur des définitions méronymiques, pour établir des listes de noms de parties du corps. La liste *morceau, partie, part*, obtenue par le biais des synonymes de *part*, puis introduite dans l'objet DEFINITION, leur a permis d'obtenir une série de noms de partie du corps – non sans avoir dû effectuer un tri. Cette série a été ensuite utilisée pour établir, puis étudier, dans le cadre d'un petit dossier de recherche, une liste de locutions contenant de tels noms – les lexicographes du *TLF* s'étant attachés à une présentation scrupuleuse des locutions[9], celles-ci sont distinguées dans l'objet SYNTAGME. Il reste encore, à partir de ces résultats, plusieurs voies de recherche ultérieures à l'étudiant, et en tout premier lieu de classification des locutions, tant les éléments rassemblés dans cet objet textuel sont divers, à différents titres (catégorie et empan de la locution, degré de figement etc.).

Troisième exemple : combiner informations morphologiques et sémantiques

Le troisième exemple rapporte l'expérience d'un dossier effectué par une étudiante sur la polysémie de noms en *-erie* : il s'agit ici d'une polysémie spécifique et supposée régulière, qui peut s'exprimer comme relevant du rapport (métonymique) entre un sémème collectif correspondant à un ensemble d'individus (humains ou objets) et un sémème locatif susceptible d'être, ou de figurer, le « contenant » de ces individus.

Cette recherche se base sur une hypothèse de départ, appuyée par quelques exemples : la polysémie qui unit deux des significations de *bijouterie* par exemple – significations décrites toutes deux dans l'article[10]. Elle exploite de manière optimale plusieurs des ressources du *TLFi* et les combine :

- recherche, par observation des définitions, ou par le biais de l'objet SYNONYME/ANTONYME, de plusieurs définissants exprimant le 'lieu' et établissement d'une liste (*lieu, endroit, place, magasin, boutique*) – liste *synlieu* ;
- recherche de définissants de noms collectifs par observation de définitions de noms collectifs et établissement d'une liste (*ensemble, groupe* ...) – liste *synensemble* ;
- établissement d'une liste de mots se terminant par *-erie* – liste *finerie* ;
- croisement des résultats de ces différentes requêtes et limitation des résultats aux substantifs.

La requête, exprimée dans le Tableau 2, donne 68 résultats : *armurerie, bijouterie, soufflerie* etc. On remarquera que l'étudiante a ici joué sur la place des définissants 'lieu' et 'ensemble' (voir les listes correspondantes), admettant dans sa requête une position allant jusqu'au 3ème mot dans l'objet DEFINITION (code &d3) : elle autorise ainsi une certaine « souplesse » dans l'expression des définitions (*petit ensemble* sera alors accessible), mais s'expose à davantage de bruit. Ici comme ailleurs, un tri manuel ultérieur est indispensable.

n° d'objet	type de l'objet	lien	contenu
1	entrée		*&lfinerie*

[9] Voir sur ce point les remarques de Imbs (1971) dans la préface du TLF.

[10] Il y a ici, bien sûr, sélection de deux sémèmes sur l'ensemble possible des sémèmes d'un substantif. Par ailleurs, l'exercice n'est pas supposé épuiser l'ensemble des valeurs du suffixe *-erie*.

2	Définition	Dépendant de l'objet *1*	*&d3&lsynlieu*
3	Définition	Dépendant de l'objet *1*	*&d3&lsynensemble*
4	Code grammatical	Inclus dans l'objet *1*	*substantif*

Tableau 2 : noms en -erie présentant une polysémie régulière lieu/ensemble des individus présents en ce lieu

Comme elle le rapporte dans son dossier, l'étudiante s'est néanmoins heurtée à divers écueils, parmi lesquels ceux concernant les « variantes notationnelles » que décrit Martin (2001 : 102). Corbin et al. (1995) évoquaient déjà la diversité des libellés sémantiques permettant de définir les noms désignant des statuts sociaux liés à une activité. Ici, c'est une diversité similaire des définissants de lieux (*établissement où l'on…*), ainsi que l'indécidabilité du classement sémantique de certains d'entre eux qui interdisent une recherche exhaustive. Comme l'illustre bien l'exemple de ce dossier, on peut néanmoins accorder à ces écueils une vertu : celle d'obliger les étudiants à passer par une recherche préalable minutieuse des variations, afin de recenser et d'expliciter, fût-ce pour eux-mêmes, les sources de cette variation. Ils doivent ainsi développer des connaissances de vocabulaire, se familiariser avec le métalangage et le métadiscours utilisés par le lexicographe pour ensuite faire des hypothèses sur les différentes manières dont un article peut être rédigé.

2. *Frantext* : initiation aux méthodes de la linguistique de corpus

C'est une recherche inductive comparable dans la base textuelle *Frantext* que nous avons également cherché à encourager chez nos étudiants. Ici, la recherche est menée dans une autre dimension, celle de textes intégraux : alors que, dans le *TLF*, l'utilisateur est confronté majoritairement à du métalangage et n'aborde les attestations que par le biais des exemples[11], il accède, avec *Frantext*, à des textes entiers, intégrés dans une interface offrant différentes fonctionnalités. *Frantext* offre donc une initiation à l'utilisation de corpus de textes, en associant la richesse d'une base textuelle organisée par genres et par périodes, et une interface de recherche combinant différentes fonctionnalités (Pierrel 2003). Le niveau d'apprentissage requis pour « entrer » dans *Frantext* est moindre. Par contre, les objectifs de recherche sont plus difficiles à appréhender par les étudiants. Dans un dictionnaire tel que le *TLFi*, les catégories à explorer sont déjà en partie calibrées par la liste d'objets textuels qu'il est possible de solliciter dans une requête. Avec *Frantext*, il est plus difficile de comprendre quel bénéfice on peut tirer d'une base de textes et en quoi peut consister l'interrogation de ce flux de données non structurées. L'enjeu de l'enseignement est alors de rendre accessibles tout à la fois les méthodes et les objectifs de la description linguistique à partir de corpus.

2.1. Découverte d'un corpus diversifié et annoté

Avec *Frantext*, la recherche passe tout d'abord par le choix d'un corpus pertinent eu égard à l'objectif, ce qui suppose, sinon des hypothèses clairement formulées, du moins une intuition préalable. Les critères de sélection au sein de ce vaste ensemble de textes sont la date, le genre (roman, théâtre, essai, mémoires…), l'auteur, le titre de l'ouvrage. Ces critères permettent par exemple d'initier des études comparées entre périodes ou genres différents.

Le corpus étant constitué, des modes d'accès différenciés sont disponibles, selon la nature des observables que l'on cherche à construire. L'interface propose en effet un mode de recherche dans les textes à partir d'un « mot » (forme, lemme[12], catégorie) ou d'une séquence de mots. Elle permet également de constituer des grammaires formelles pour décrire des contextes

[11] A la différence d'autres dictionnaires, les exemples du TLF sont tous attestés ; ils sont parfois d'une longueur considérable, ce qui constitue une autre spécificité du TLF.

[12] Un lemme (forme de base) est une forme lexicale existante choisie par convention au sein d'un paradigme flexionnel : forme de l'infinitif pour le verbe, du masculin singulier pour l'adjectif par exemple.

complexes. Elle offre enfin un module qui permet de visualiser de façon plus synthétique les principaux cooccurrents d'un mot (module de VOISINAGE). Le travail initié dans le cadre du *TLFi* se prolonge : il s'agissait dans le cas du dictionnaire de déterminer quels objets permettent d'accéder à l'information recherchée. Avec *Frantext*, il faut choisir les bonnes catégories de description et le bon instrument d'observation.

Découvrir le travail sur un corpus de textes, c'est également comprendre d'emblée que les données peuvent être enrichies d'annotations qui permettent d'interroger non seulement les formes d'occurrences mais les lemmes, voire les catégories grammaticales. La version catégorisée de *Frantext* a été traitée par un programme qui découpe le texte en entités et leur assigne une catégorie grammaticale qui peut être intégrée dans la requête. On accède alors à des recherches plus générales (travail sur des patrons associant matériel lexical et informations syntaxiques) ou mieux ciblées (travail sur des formes désambiguïsées en précisant la catégorie grammaticale dont elles relèvent).

Les résultats des requêtes, présentés dans des empans de contextes relativement étroits, sont rapprochés sur la base de la pertinence des critères d'interrogation choisis. C'est principalement ce rapprochement des résultats qui constitue la première étape de la confrontation des hypothèses et de l'observation.

2.2. Un instrument pour travailler sur les phénomènes idiomatiques

La base *Frantext* permet en particulier de travailler sur la diversité des « phénomènes de l'expression idiomatique » (Legallois et François 2006), depuis les expressions figées et les locutions jusqu'aux associations plus lâches caractéristiques des mécanismes de la collocation. La possibilité d'associer catégories grammaticales et formes lexicales permet en particulier d'étudier les phénomènes collocationnels qui associent un ancrage lexical spécifique et des propriétés syntaxiques (colligations et contructions). Le fait de pouvoir faire varier l'empan du contexte de l'association (depuis les séquences contiguës jusqu'à des associations dans une fenêtre de plusieurs phrases) est également un paramètre intéressant pour étudier tous types de rapports collocatifs. Les exemples que nous présentons font varier ces paramètres et illustrent différents phénomènes idiomatiques.

Premier exemple : dimension diachronique des expressions du français

Ce premier exemple montre un cas simple d'étude à partir de la base non catégorisée (environ 4000 textes). Il vise à faire prendre conscience à l'étudiant de la dimension temporelle du corpus. Si la grande majorité des textes se répartissent sur les XIXème et XXème siècles, les périodes plus anciennes sont néanmoins bien représentées, et ont fait l'objet de mises à jour régulières, ce qui permet de faire des observations sur une large échelle de temps (du début du XVIème siècle jusqu'à aujourd'hui). On peut alors mener des études ponctuelles pour observer au fil du temps l'émergence ou les changements d'emploi de locutions ou d'expressions figées dans une perspective diachronique. Une recherche sur la séquence *à couvert* permet ainsi de constater très clairement qu'elle est exclusivement employée en tant que locution prépositionnelle dans les premiers textes (*à couvert de l'envie, de l'oppression, d'un fléau …*) alors qu'en fin de période elle n'est plus employée que comme adverbe (*cheminer, se mettre à couvert*). Il est dans ce cas très intéressant d'amener les étudiants à confronter leurs trouvailles à la description qu'en fait le TLF, dont voici un extrait :

C. — *Loc. adv.* et *prép.*
1. *Loc. adv. À couvert.* Sous la protection matérielle de quelque chose qui couvre. *Arriver à couvert, s'exercer à couvert, tirer à couvert.* […] — *Au fig.* Sous la garantie, sous la protection de. *Agir à couvert, être à couvert, se mettre à couvert, mettre qqc. à couvert.* […]
2. *Loc. prép.*

a) *À couvert de.* À l'abri de. *À couvert de la pluie, du vent, de l'orage. Se mettre à couvert du mauvais temps du dehors* (SAINTE-BEUVE, *Volupté,* t. 2, 1834, p. 191).
—*Au fig.* Sous la garantie de, à l'abri de :

La référence à l'article les aide à élaborer leurs catégories de description (les mentions *loc. adv.* et *loc. prép.* leur sont données dans l'objet CODE GRAMMATICAL). Par ailleurs, ce travail de confrontation leur montre que la dimension temporelle, qui est absente du dictionnaire peut être récupérée par une étude en corpus ; ils comprennent alors comment l'étude de corpus peut venir affiner des descriptions existantes.

Autre exemple, l'étude de l'expression *chanter les louanges*, à condition d'être correctement menée – c'est-à-dire à condition de formuler une recherche qui anticipe le phénomène de variation[13] – montre à l'œuvre le mécanisme de figement, la stabilité de la locution en fin de période contrastant avec des occurrences initiales très variées (inversion, variation en nombre du mot *louanges* : *qui par eux votre louange chante*, absence de déterminant : *chante louanges*, coordination : *chantans hymne et louanges*). Pour des étudiants qui ne sont que trop rarement familiarisés avec la dimension diachronique, ce type d'activité a le double mérite de leur faire découvrir la réalité d'un état passé de leur langue et de les amener à réfléchir aux phénomènes de variation dans le temps.

Deuxième exemple : étude de phénomènes de collocation

Si l'on combine cette fois des informations lexicales et syntaxiques, la base étiquetée offre la possibilité d'associer des formes lexicales et des traits syntaxiques dans la formulation de la recherche. On peut s'intéresser par exemple à des affinités entre verbe et adverbe. Un exercice proposé consistait à déterminer les verbes auxquels s'associe le plus souvent l'adverbe *nerveusement*. Une première option de recherche consiste à extraire tous les contextes comprenant une occurrence du patron : verbe + *nerveusement*. On consulte la base étiquetée à partir de la requête suivante, où « V » est la codification de la catégorie Verbe :

&e(g=V) nerveusement

Dans ce cas, on récupère une série de 191 contextes à partir desquels il est possible de dégager la distribution de l'adverbe. Une deuxième option, qui constitue une alternative intéressante au balayage de ce grand nombre de contextes accompagné du relevé et du comptage de toutes les instances de la catégorie verbe, consiste à utiliser le module VOISINAGE pour calculer les principaux cooccurrents de l'adverbe. On définit pour cela la taille de l'empan de texte à considérer (dans notre cas, 0 mot avant la forme, 1 mot après), et on trie les cooccurrents par fréquence décroissante. Les premières formes obtenues sont : *parlant*, *rit*, *serrait*, *rire*, *serrant*, *agitait*, *battaient*, *marchait*, … Les étudiants sont alors amenés à découvrir la complémentarité d'outils de comptage et de calcul de cooccurrence, qui offrent une vue synthétique du comportement de certaines unités dans le corpus considéré, et d'outils d'exploration contextuelle, qui facilitent l'interprétation de ces résultats par l'examen attentif d'occurrences particulières.

Troisième exemple : vers l'étude des constructions

L'exploitation de l'étiquetage permet d'accéder à l'examen de constructions particulières. Un exercice consiste à partir d'expressions connues – *plus de peur que de mal*, *plus de mal que de bien* – pour examiner le caractère productif de la construction correspondante en retrouvant d'autres expressions formées sur ce modèle : on cherche ici des expressions mettant en

[13] Le plus simple est de rechercher les extraits dans lesquels le lemme *chanter (&cchanter*) et le lemme *louange* (*&mlouange*) apparaissent ensemble dans un contexte proche (par exemple, séparés au maximum de 3 mots), dans un ordre indifférent.

parallèle deux substantifs sémantiquement apparentés (antonymes ou co-hyponymes). Le travail commence alors par un petit exercice d'abstraction, qui consiste à dégager un patron lexico-syntaxique à partir de ces deux exemples, ce qui donne :

plus de &e(g=S) que de &e(g=S)

Voici un extrait des résultats affichés :

(i) Une taille svelte et bien prise annonçait *plus de légèreté que de vigueur.*
(ii) Jeanne parlait avec *plus de vivacité que de coutume*.
(iii) C'était une poupée de Paris […] avec *plus de charme que de beauté.*
(iv) Il n' y a pas *plus de défaite que de beurre* aux fesses
(v) […] après une revue de chemises qui avait mis en évidence *plus de loques que de linge décent*

On le voit, tous les résultats ne sont pas pertinents. Cet exercice demande donc aux étudiants de se livrer à un examen scrupuleux des résultats de façon à sélectionner parmi des résultats bruités les instances véritables de la construction étudiée : ils doivent être à même de comprendre que certaines occurrences ne sont pas conformes à l'objet recherché, alors même qu'elles correspondent au patron de recherche. Ils doivent également être en mesure d'en comprendre la raison. Ce travail d'interprétation est particulièrement formateur, parce qu'il sollicite les connaissances linguistiques des étudiants pour démêler les bonnes et les mauvaises instances du patron qu'ils ont conçu. Les résultats à éliminer peuvent être imputables au processus d'étiquetage. Dans ce cas, ce sont des connaissances grammaticales élémentaires qui doivent être activées. En (ii), l'adverbe *de coutume* n'ayant pas été identifié par le programme, il a été mal segmenté et étiqueté et se trouve analysé comme une séquence [de + nom]. Il est plus difficile de voir qu'en (iv) c'est une autre construction qui prend le relais (pas plus de X que de Y), ce qui explique que les deux noms ne puissent être considérés comme des co-hyponymes… On peut alors amener l'étudiant à réfléchir à des moyens de raffiner la requête, par exemple en éliminant de la façon suivante les contextes de négation (élimination de la forme *pas*) :

^pas plus de &e(g=S) que de &e(g=S)

On amorce ici la démarche décrite par Leroy (2004) comme un « aller-retour entre analyse linguistique et repérage informatique » : la mise au point des patrons de recherche vient affiner la connaissance du phénomène qu'il s'agit de décrire, dans la mesure où l'identification de résultats proches de ceux visés aide à circonscrire les limites de la catégorie que l'on étudie.

Un dernier exemple, développé dans Frantext catégorisé, montre que l'on peut encore s'abstraire du niveau lexical dans l'étude des phénomènes constructionnels. Nous appuyant sur une description issue de la *Grammaire Méthodique du Français* (p.165), nous proposons aux étudiants l'objectif suivant :

> « Un grand nombre d'expressions verbales lexicalisées ou quasi lexicalisées contiennent un complément d'objet sans déterminant. Ex : *faire long feu*, *rendre hommage*. Formulez une requête permettant de les repérer dans *Frantext* (éliminez les verbes *être* et *avoir*). Relevez les 20 premières occurrences. Triez-les en conservant celles qui correspondent effectivement à des expressions verbales lexicalisées. Expliquez pourquoi vous ne retenez pas les autres, autrement dit, spécifiez la source de bruit. ».

Après différents tâtonnements, les étudiants sont amenés à proposer deux requêtes, la première pour identifier les expressions comportant seulement un nom, la seconde pour prévoir la présence d'un adjectif en position prénominale :

&e(g=V c!=(&cavoir|&cêtre)) &e(g=S)

&e(g=V c !=(&cavoir|&cêtre)) &e(g=A) &e(g=S)

Voici quelques exemples de résultats pour les deux requêtes :

verbe + nom	Verbe + adjectif + nom
prend forme	*laissaient pleine liberté*
donneront naissance	*jouaient franc jeu*
devenait corsaire	*prenaient grand soin*
rendît responsables	*fane jeune fille*
demanda compte	*donnent seules ouverture*

Tableau 3 : résultats de deux requêtes recherchant des locutions verbales

Cette fois, la confrontation avec les résultats doit amener l'étudiant à faire la différence entre cette construction spécifique et des constructions attributives (*devenait corsaire*) ou à repérer la présence de sujets inversés (*fane jeune fille*); la présence d'intrus s'explique également par le fait qu'il n'est pas possible de s'assurer que l'adjectif et le nom sont liés syntaxiquement, d'où la présence de séquences comme *donnent seules ouverture à*, qu'il ne s'agit pas d'éliminer de la recherche mais d'analyser pour en extraire la locution *donner ouverture* en considérant *seules* comme une incise. Enfin, une des erreurs classiques de l'étiquetage automatique, que l'on retrouve dans Frantext, provient de la difficulté à discriminer les emplois adjectivaux et nominaux (*rendît responsables*). Malgré ces sources d'erreur, présentes en amont de la recherche elle-même, on trouve là un moyen efficace d'identifier les locutions considérées : il s'agit en effet d'une méthode qui offre un bon compromis entre la simplicité de mise au point du patron et le coût de sélection manuelle.

En situation d'enseignement, il est indispensable de trouver des objets de recherche qui se prêtent facilement à ce type d'approche, c'est-à-dire qui ne génèrent pas un travail d'analyse et de tri trop conséquent en aval de la recherche. C'est la raison pour laquelle il est difficile avec *Frantext* de proposer aux étudiants des activités peu cadrées, au cours desquelles ils seraient amenés à concevoir de façon autonome des objets de recherche propres.

3. Prolonger et généraliser la démarche outillée

Le travail à partir du *TLFi* et de *Frantext* constitue selon nous une étape extrêmement utile dans la formation d'étudiants de linguistique. Ils offrent pour l'enseignant comme pour l'étudiant un environnement favorable pour initier le travail d'investigation de données linguistiques et pour ancrer la description linguistique dans une démarche empirique qui débute par la constitution d'hypothèses de recherche et la collecte d'observables pertinents. Néanmoins, cet apprentissage doit être considéré comme une première étape dans un processus d'apprentissage plus complet des méthodes et des outils qui relèvent plus généralement d'une approche outillée de la linguistique (Habert 2005). Plusieurs aspects viennent en effet limiter les possibilités de mener à partir de ces outils des recherches plus approfondies et de plus grande ampleur.

Dans le cas du *TLFi*, les étudiants découvrent une base de données lexicales extrêmement riche et totalement libre d'accès, dont on peut donc espérer qu'elle sera désormais pour eux une référence qu'ils consulteront régulièrement. On regrette néanmoins que cette formidable ressource ne soit pas mise à jour. Il est en effet impossible de travailler avec le *TLFi* sur des questions concernant l'état actuel du lexique ou les processus de néologie, qui sont des problématiques attractives pour les étudiants qui débutent en linguistique. Par ailleurs, les potentialités de recherche sont contraintes par l'écueil des variations notationnelles, et Martin (2001) a bien montré l'écart entre un dictionnaire informatisé de ce type et une base de

données lexicale se prêtant à une exploitation automatisée. En outre, les recherches présentant une dimension sémantique sont rapidement limitées par la difficulté d'accéder à une information de cette nature.

Concernant *Frantext*, les limites d'exploitation de ce matériau très riche sont liées à la nature du corpus lui-même. Il faut en effet se souvenir que *Frantext* a été d'abord rassemblé pour servir un projet lexicographique, et que la sélection des textes à l'origine s'est faite en fonction des possibilités de l'époque (d'où en particulier l'absence de composante orale) et d'une certaine idée du corpus idéal pour le dictionnaire envisagé (reflétant une « culture de type humaniste actualisée », selon la formule de Imbs). Malgré des mises à jour régulière, le corpus reste marqué par cette orientation initiale, et n'offre pas les qualités de représentativité d'un corpus équilibré et échantillonné tel que le *British National Corpus*. En outre, la présentation des résultats en contexte manque de souplesse en comparaison des techniques disponibles dans les concordanciers librement accessibles à l'heure actuelle, qui offrent des modalités indispensables de tri, de réglage de l'empan de texte, et de recherche.

En conséquence, le travail réalisé à partir de ces ressources doit être considéré comme un tremplin vers la découverte d'autres ressources et d'autres outils d'exploration. La situation idéale pour lancer un travail plus approfondi consiste en effet à laisser à l'étudiant plus d'initiative dans la mise en place de son dispositif d'exploration, en commençant par la constitution de son propre corpus et le choix d'instruments d'analyse adaptés pour élaborer la sélection d'observables qu'il juge intéressants. Cela suppose une diversification des apprentissages, et idéalement une découverte de la programmation (cf. Leroy 2004). C'est un travail qui est mené également à Toulouse, parallèlement à celui que nous avons évoqué ici, et qui vise à doter les étudiants de compétences en linguistique de corpus qui sont encore insuffisamment intégrées dans les cursus de sciences du langage en France.

4. Remerciements

Nous remercions Nabil Hathout qui nous a fait bénéficier de son expérience d'enseignement de *Frantext* à l'Université de Bordeaux et Josette Rebeyrolle, dont les remarques nous ont permis d'améliorer la présentation d'un enseignement auquel elle a elle-même collaboré.

Bibliographie

Corbin D., Corbin P., Tutin A., Aliquot S. (1995) : « Ce que des linguistes peuvent attendre d'un dictionnaire informatisé », in D. Piotrowski (éd.), *Lexicographie et informatique. Autour de l'informatisation du Trésor de la langue française*, Actes du colloque international de Nancy (29, 30, 31 mai 1995), Paris, Didier Erudition, p. 51-77.

Dendien J., Pierrel J.-M. (2003) : « Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence », *Revue TAL (Traitement Automatique des Langues)*, numéro sur les dictionnaires électroniques, Hermes Sciences Edition, vol. 44, n° 2, p. 11-38.

Legallois, D., François, J. (2006) : « Autour des grammaires de construction et de patterns » Cahier du CRISCO n° 21, Université de Caen.

Habert B. (2005) : *Instruments et ressources électroniques pour le français*, Paris/Gap : Ophrys (« L'essentiel français »).

Imbs P. (1971). « Préface du Trésor de la Langue Française » [en ligne] http://www.atilf.fr/atilf/divers/Preface_TLF.html

Leroy S. (2004) : « Extraire sur patrons : allers et retours entre analyse linguistique et repérage automatique », *Revue française de linguistique appliquée*, IX-1, 25-43.

Martin R. (2002) : *Comprendre la linguistique*. Presses Universitaires de France, Paris.

Martin R. (2001) : *Sémantique et automate*, *l'apport du dictionnaire informatisé*, Écritures électroniques, Presses Universitaires de France, Paris.

Pierrel J.-M. (2003) : « Un ensemble de ressources de référence pour l'étude du français : *TLFi*, *Frantext* et le logiciel Stella », *Revue québécoise de linguistique*, numéro sur TALN, Web et corpus, vol.32, n°1, p. 155-176.

Pruvost J. (2006) : *Les dictionnaires français outils d'une langue et d'une culture,* L'Essentiel français, Ophrys.

Rastier F. (2005) : « Enjeux épistémologiques de la linguistique de corpus », in G. Williams (éd.), *La linguistique de corpus, Actes des 2èmes journées Linguistique de Corpus à Lorient*, PU Rennes, Rennes, p. 31-45.