Exploitation linguistique de corpus arborés d'écrits scientifiques, à l'aide du logiciel ScienQuest

Achille Falaise, laboratoire LIG-GETALP

1 Introduction

Les travaux linguistiques présentés dans cet ouvrage, et menés dans le cadre du projet Scientext, s'appuient sur plusieurs corpus de textes scientifiques, annotés linguistiquement. Ce chapitre présente le matériau de base de ces travaux (les corpus et leur annotations) ainsi que les outils développés pour son exploitation.

Les corpus du projet Scientext sont enrichis d'annotations, typologiques, thématiques, lexicales et syntaxiques (on parle de corpus arborés). De plus en plus de corpus comportent ces types d'annotations. Le fait qu'elles soient présentes directement dans le corpus simplifie le travail des linguistes, qui peuvent ainsi tester et valider rapidement leurs hypothèses directement à ces niveaux d'abstraction.

Encore faut-il que ces annotations soient exploitables. À la base, celles-ci sont présentes sous forme de métadonnées au sein des corpus, qui sont accessibles à des programmes informatiques, mais ne sont pas destinées à être exploitées directement par des utilisateurs non initiés à la linguistique-informatique. Ce problème n'est pas trivial, et les outils actuels d'exploration de corpus annotés sont souvent complexes à utiliser pour un utilisateur non-spécialiste. L'ergonomie et la facilité d'utilisation de ces outils restent donc aujourd'hui un enjeu majeur. C'est ce qui nous a amené à proposer ScienQuest, un environnement de recherche simple, adapté aux linguistes, didacticiens, lexicographes ou épistémologues, pour l'étude de corpus richement annotés.

Nous présenterons dans un premier temps les corpus du projet Scientext, et leurs annotations. Nous définirons ensuite la notion de « convivialité » appliquée aux outils d'exploitation de corpus, et montrerons, à travers l'étude de quelques exemples tirés des travaux présentés dans cet ouvrage, en quoi ScienQuest simplifie l'exploitation de corpus annotés. Nous conclurons sur un bilan de l'utilisation de cet outil, dans le cadre du projet Scientext mais aussi de sa réutilisation dans d'autres projets.

2 Les corpus du projet Scientext

Dans le cadre du projet Scientext, nous avons rassemblé quatre corpus distincts, en français et en anglais, couvrant différents aspects de l'écrit scientifique. Ces corpus sont librement consultables dans ScienQuest. Les 4 corpus de référence du projet Scientext sont les suivants :

- un corpus de 219 **textes scientifiques en français** (4,8 millions de mots),
- un corpus de 3381 **articles de biologie et de médecine en anglais** (13,8 millions de mots),
- un corpus de 300 **textes argumentatifs d'apprenants de l'anglais** (1,1 millions de mots),
- un corpus de 502 **évaluations de communications** du colloque CÉDIL[1].

2.1 Structure et annotations structurelles

Chacun des corpus se compose de textes, et chaque texte est caractérisé par des annotations structurales internes : par exemple son type (thèse, article scientifique, etc.), sa discipline, etc. D'autre part, à l'intérieur de chaque texte, on distingue différentes parties textuelles, par exemple l'introduction, la conclusion, etc. Cela permet de mener des études portant uniquement sur des

1 CÉDIL : Colloque international des Étudiants chercheurs en Didactique des Langues et en Linguistique.

textes comportant certaines caractéristiques (par exemple uniquement sur des thèses de linguistique), mais surtout de mener des études contrastives ou distributionnelles (par exemple, on peut comparer le nombre et la forme des prises de positions entre les introductions et les conclusions). Le détail des annotations varie suivant les corpus :

- **Textes scientifiques en français** : les textes de ce corpus couvrent huit disciplines (linguistique, psychologie, sciences de l'éducation, traitement automatique des langues, biologie, médecine, électronique et mécanique) et quatre types de textes (articles, communications, mémoires de thèse et d'habilitation à diriger des recherches), et sont décomposés en huit parties textuelles (développement, introduction, conclusion, résumé, notes, titres, remerciements et annexes).
- **Articles de biologie et de médecine en anglais** : les textes de ces corpus sont décomposés en cinq parties textuelles (développement, introduction, conclusion, résumé et titre). Un travail est actuellement en cours pour distinguer ces textes en fonction de leur discipline, et affiner la décomposition en parties textuelles.
- **Textes argumentatifs d'apprenants de l'anglais** : on distingue les textes en fonction du niveau des étudiants (deuxième ou troisième année), et les textes sont découpés en quatre parties textuelles (développement, introduction, conclusion et titre).
- **Évaluations de communications** : on distingue dans ce corpus les textes en fonction de quatre disciplines (description linguistique, psycholinguistique et développement langagier, sociolinguistique et plurilinguisme, et didactique des langues), du destinataire (messages pour les auteurs, ou bien pour les organisateurs), et de l'avis (avis d'acceptation, de rejet, ou réservé).

Ces annotations ont été effectuées manuellement, suivant le standard XML TEI Lite.

2.2 Annotations morphosyntaxiques

Outre des annotations d'ordre structural, les corpus de Scientext comportent des annotations morphosyntaxiques. Chaque phrase est ainsi décomposée en lexèmes (mots, ou expressions figées). Pour chacun de ces lexèmes, sont précisés des annotations morphologiques : la forme (c'est à dire le mot effectivement trouvé dans le texte), le lemme (c'est à dire le mot non fléchi, tel qu'on le trouverait dans une entrée de dictionnaire), la partie du discours (catégorie syntaxique : nom, verbe, etc.) et des traits flexionnels (genre, nombre, participe passé/présent, etc.).

Par exemple, la forme « collocations » est analysée comme suit :

- forme : *collocations* ;
- lemme : *collocation* ;
- partie du discours : *nom* ;
- flexion : *genre indéterminé*, *pluriel*.

Pour cet exemple, nous savons que « collocation » est un mot féminin en français, mais le système d'annotation n'a pas pu le déterminer automatiquement, c'est pourquoi le corpus contient l'annotation « genre indéterminé ».

Au delà de la dimension morphologique, le corpus comporte aussi des annotations syntaxiques en dépendances, c'est à dire les types de relations syntaxiques qui lient les lexèmes. Cela permet de dépasser le niveau de la séquence de mots, pour effectuer des recherches sur les relations syntaxiques entre mots.

Un exemple complet d'analyse est donné en figure 1.

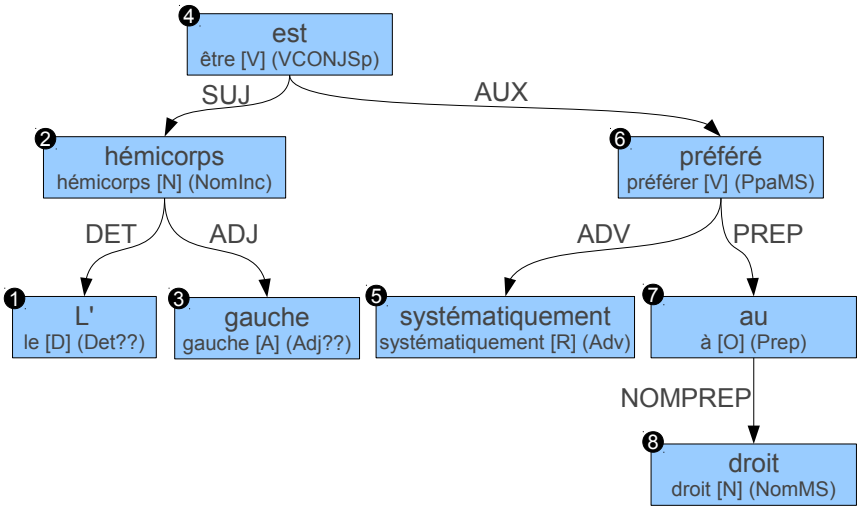

Figure 1: analyse morphosyntaxique avec Syntex de l'énoncé « L'hémicorps gauche est préféré systématiquement au droit ».

Le coût d'une annotation morphosyntaxique manuelle de qualité étant extrêmement élevé, il n'était pas envisageable d'annoter ainsi manuellement toutes les phrases des corpus. C'est pourquoi, contrairement aux annotations structurales, les annotations morphosyntaxiques ont été obtenues automatiquement à l'aide de Syntex, un logiciel d'analyse automatique de textes (Didier Bourigault, 2007). L'annotation obtenue n'est par conséquent pas exempte d'erreurs. Par exemple, le système dispose d'heuristiques basées sur le contexte pour désambiguïser les formes textuelles comme *est*, qui peuvent correspondre soit au verbe *être*, soit au nom commun *est* (le point cardinal), mais des erreurs subsistent. Cela est particulièrement sensible en anglais, où la simplicité de la morphologie entraîne de nombreuses ambiguïtés, particulièrement difficiles à résoudre lorsque les formes du contexte souffrent elles-mêmes d'ambiguïtés.

2.3 Droits d'accès aux corpus

Chaque corpus de Scientext existe en deux versions, la première étant uniquement téléchargeable et la seconde uniquement consultable dans ScienQuest :

1. Une **version sans annotation morphosyntaxique, mais avec annotation structurale**, téléchargeable sous licence *Creative Commons*[2]. Cette version comprend les annotations structurelles (types de textes, disciplines, parties textuelles, etc.), mais pas les annotations morphosyntaxiques (lemmes, traits morphologiques, relations syntaxiques).
2. Une version **avec annotation morphosyntaxique *et* structurale**, consultable dans ScienQuest, mais non téléchargeable, la licence de l'analyseur Syntex ne le permettant pas.

3 Outillage de corpus pour l'utilisation en linguistique

La façon d'outiller un corpus dépend du public et du type d'étude visés. Dans le cas de Scientext, il s'agit d'un public de linguistes, didacticiens, lexicographes et épistémologues non-informaticiens, pour des tâches de recherche de motifs linguistiques dans un corpus ou un sous-corpus.

2 http://creativecommons.org/

3.1 Enquête et scénario d'utilisation

Lors de l'élaboration d'un outil, il est important de savoir si l'outil que l'on conçoit est utilisable par le public visé. Un critère nous semble particulièrement important : l'expertise informatique de l'utilisateur. En effet les outils d'exploitation de corpus sont généralement créés par des TAListes, c'est à dire des utilisateurs ayant un bon niveau d'expertise informatique, qui n'ont pas nécessairement conscience des difficultés d'un utilisateur peu expert.

Une première enquête a eu lieu en mars 2008 auprès d'une quinzaine de chercheurs et d'étudiants en linguistique, didactique et communication, issus de plusieurs laboratoires de recherche (LIDILEM, LLS, LiCoRN et GRESEC). Il s'agissait de réfléchir à l'élaboration d'une interface permettant d'interroger un premier échantillon du corpus Scientext constitué d'écrits scientifiques analysés syntaxiquement et structurellement. Plusieurs utilisateurs ont clairement indiqué leur difficulté à utiliser les outils disponibles alors, y compris des outils assez classiques comme Frantext, pour lequel la recherche à l'aide d'expressions régulières est difficilement maîtrisée par des utilisateurs occasionnels. Nous leur avons demandé d'exprimer à l'aide d'exemples leurs besoins en termes de recherches dans les corpus du projet Scientext, puis de se prononcer sur une première maquette d'interface, et en particulier sur son accessibilité pour des utilisateurs peu experts en informatique.

En dialoguant avec les utilisateurs, un scénario générique a pu être mise en place, qui se compose de trois étapes :

1. Définition d'un sous-corpus à partir des textes du corpus, en fonction des disciplines, types de textes, parties textuelles, etc. définies dans le corpus.
2. Définition d'une thématique ou d'un motif de recherche que l'utilisateur souhaite étudier.
3. Présentation des résultats, à la fois la liste en contexte des occurrences trouvées dans le sous-corpus, et des statistiques sur leur répartition dans le sous-corpus. L'exportation de ces résultats vers des tableurs, fréquemment utilisés par les linguistes, doit être possible.

3.2 Convivialité des outils pour des non informaticiens

Suite à ces consultations, nous avons pu établir trois exigences pour définir un environnement de recherche sur corpus convivial, facilement utilisable par des non informaticiens :

- **Absence de technicité.** Il doit être utilisable sans connaissance préalable, en tout cas pour une première approche, d'un langage de requête spécifique ou d'un langage de balisage comme XML. Les éléments spécifiques ou techniques devront être transformés par des valeurs pré-établies intégrées dans des ascenseurs ou des listes à cocher. Les termes employés devront être le moins techniques possible, ce qui constitue un véritable défi pour des annotations linguistiques complexes.
- **Rapidité et facilité d'emploi.** Le système doit être rapide et simple d'emploi. L'utilisateur ne doit pas avoir à parcourir de documentation, en tout cas pour une utilisation standard. L'usager sera guidé dans sa démarche tout au long du processus.
- **Expressivité et progressivité.** Le mode assisté doit permettre d'exploiter le mieux possible la richesse de l'annotation. Il est intéressant de prévoir une progressivité d'un mode simple à un mode plus complexe, dans une démarche didactique. Il est évidemment impossible de proposer en mode simple assisté toute la richesse qu'offre un langage de requêtes complexe. Il sera néanmoins intéressant d'amener l'utilisateur « en douceur » à cette progressivité.

Peu d'outils tentent de répondre à ces exigences ; ils sont souvent basés sur des langages de requête formels, qui les destinent à des utilisateurs experts. Par exemple, ConcQuest (Kraif 2008) utilise un langage formel spécifique pour décrire les motifs de recherche. Des logiciels graphiques existent, en particulier le logiciel TigerSearch (Lezius 2002, Voormann 2002). Toutefois, l'éditeur, qui consiste essentiellement en un éditeur d'arbres annotés, reste très proche du langage de requête formel sous-jacent. Il simplifie la tâche d'un utilisateur expérimenté, mais reste très complexe pour un novice.

Notre approche s'inscrit dans la lignée des quelques rares outils conçus pour être pleinement

utilisables par des non informaticiens, comme l'interface du *Russian National Corpus*[3] et l'interface du corpus étiqueté et lemmatisé Elicop[4] (Mertens 2002).

3.3 Conception d'un outil pour l'étude linguistique de textes par des non informaticiens : ConcQuest

À la suite de la première enquête auprès d'utilisateurs, un prototype a été élaboré puis évalué. Cette étape a été itérée plusieurs fois[5].

Une première évaluation (et mise à disposition) a notamment eu lieu sur la version 0.8 de ScienQuest. Cette version comportait la sélection de sous-corpus, la recherche en mode avancé (par langage de requête), l'affichage KWIC des résultats, et le calcul de statistiques ; les recherches sémantiques (prédéfinies) et libres (à l'aide d'un assistant) n'étaient pas encore disponibles (voir partie 4 pour la description détaillée des fonctionnalités de ScienQuest). Certains utilisateurs, ayant du mal à formuler des besoins en termes formels, ont demandé à travailler sur les collocations, en lien avec les thèmes linguistiques du projet. Ces premières évaluations ont montré l'intérêt de la sélection de sous-corpus par critères, et du calcul de statistiques, mais le langage de requête était toujours jugé trop complexe par les utilisateurs, qui se cantonnaient généralement à des recherches très simples du type cooccurrence de deux lemmes contigus.

Une deuxième évaluation, intégrant un assistant pour un mode de recherche simple, a été proposée auprès de chercheurs internes au projet (LIDILEM, LLS) et externes (CECL[6] de Louvain). Ce mode de recherche libre et assisté a été accueilli très favorablement. En particulier, beaucoup d'utilisateurs ont alors commencé à travailler avec les relations syntaxiques à partir de cette version (notamment pour l'extraction de la phraséologie), alors que peu d'entre eux utilisaient les relations syntaxiques avec le langage de requête du mode avancé jugé trop complexe. Ainsi, à l'aide du mode simple, les utilisateurs ont pu mieux exploiter les possibilités offertes par les annotations du corpus. Des besoins supplémentaires sont apparus, comme le traitement de la syntaxe profonde (en particulier les passifs dont l'analyse avec Syntex n'est pas très intuitive) ou des recherches portant sur la ponctuation et non seulement les mots ; l'émergence de ces besoins témoigne d'une meilleure prise en main par les utilisateurs.

Ce cycle de développement « en spirale » de l'interface, en lien avec les retours des utilisateurs, a été renouvelé jusqu'à la version actuelle.

4 Exploitation de corpus avec l'outil ScienQuest

ScienQuest est une plateforme Web pour la consultation de corpus en ligne, développée en PHP. Elle a été initialement développée pour les corpus du projet Scientext, mais est aussi utilisée pour d'autres corpus, notamment ceux du projet Emolex[7].

4.1 Organisation de l'interface de ScienQuest

L'interface est bâtie selon le scénario établi avec notre échantillon d'utilisateurs. Nous avons fait le choix d'une approche segmentée en tâches simples et ordonnées, afin de guider l'utilisateur tout au long de l'utilisation de l'outil. L'objectif est une interface utilisable sans mode d'emploi, du moins pour une utilisation simple, par un public non-informaticien.

Ce choix est clairement visible dans l'interface. Le menu (à gauche de l'interface) détaille ainsi les trois grandes étapes que nous avons définies : définition d'un sous-corpus, recherche dans les textes et exploitation des résultats. Chacune de ces étapes est décomposée en sous-tâches plus simples, correspondant à une page de l'interface.

3 http://ruscorpora.ru/

4 http://bach.arts.kuleuven.be/elicop/

5 La première version stable de ScienQuest était la 0.9, il est maintenant en version 1.4.

6 *Centre for English Corpus Linguistics*, Université catholique de Louvain.

7 http://www.emolex.eu/

Sur chaque page, un en-tête précise le rôle de cette page et la nature de l'étape suivante. Un bouton permet d'accéder à cette étape. Ce cheminement n'est pas obligatoire : un utilisateur avancé peut cliquer sur la tâche de son choix pour accéder directement à la page en correspondante.

En bas de chaque page, il est possible de sauvegarder le contenu, ou à l'opposé d'importer un contenu préalablement sauvegardé.

Enfin l'interface est totalement multilingue, la langue peut donc être modifiée à tout moment. Actuellement, le français et l'anglais sont supportés.

4.2 Création de sous-corpus

La première étape de notre scénario consiste à déterminer sur quels textes et quelles parties textuelles on souhaite travailler. Par défaut, c'est l'ensemble du corpus qui est sélectionné.

Le premier écran de ScienQuest (figure 2) permet donc la sélection d'un sous-corpus, en fonction des critères d'organisation propres à chaque corpus. Par exemple, pour le corpus de textes scientifiques français du projet Scientext, selon la discipline, les types de documents et les parties textuelles.

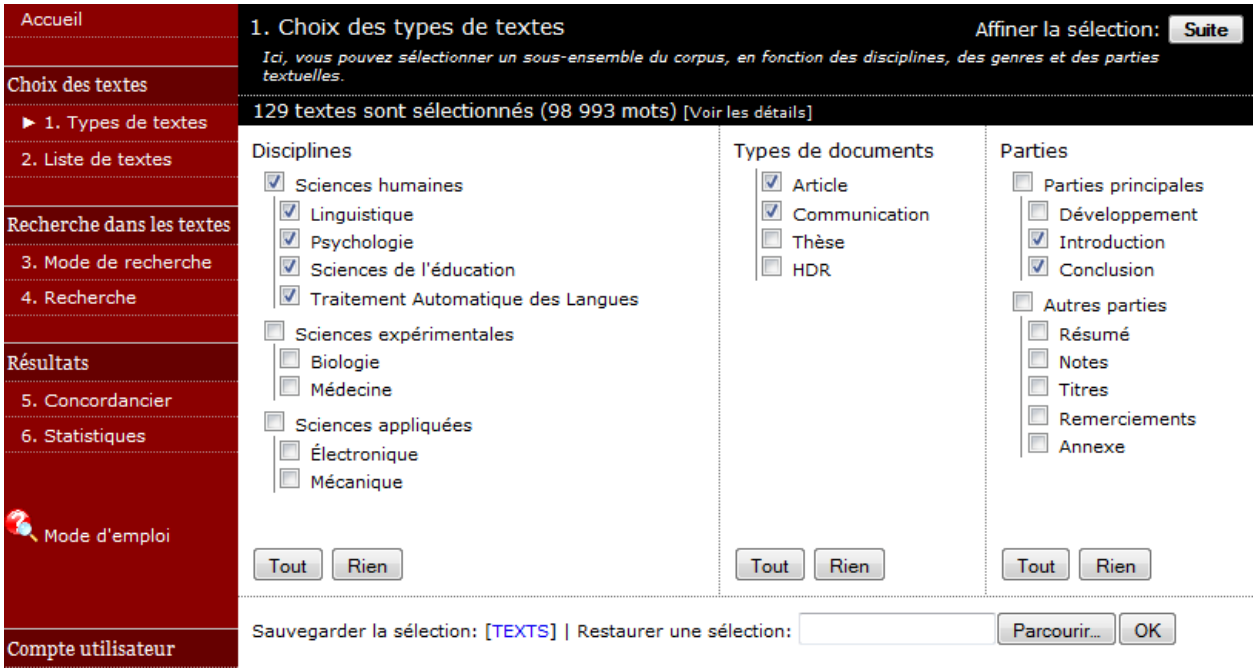

Figure 2: exemple de sélection d'un sous-corpus dans ScienQuest pour le corpus français de Scientext: les introductions et conclusion des articles de recherche et communications en sciences humaines.

Une deuxième page permet une sélection plus fine, texte par texte.

L'interface indique, pour chaque sous-corpus, le nombre de textes et de mots sélectionnés. Une quantification détaillée est disponible, de façon à faciliter la construction de corpus équilibrés. La notion de « corpus équilibré » varie suivant l'étude envisagée. Ainsi, pour son étude de la citation positionnée dans l'écrit scientifique (publiée dans le présent ouvrage), Magda Florez a utilisé cette fonctionnalité avancée pour constituer un sous-corpus homogène de textes de linguistique, psychologie et sciences de l'éducation, comportant pour chaque discipline 50 articles et 5 thèses. De leur côté, Monika Bak Sienkiewicz et Iva Novakova, dans leur étude du raisonnement causal dans les textes scientifiques (elle aussi publiée dans le présent ouvrage), ont préféré un sous-corpus homogène en termes de mots par discipline (450k mots par discipline).

4.3 Recherche dans le corpus

Une fois le corpus sélectionné selon les disciplines, les genres textuels et les parties textuelles

désirés, l'utilisateur peut effectuer une recherche. C'est surtout la recherche qui peut bloquer un utilisateur non-expert, c'est pourquoi nous avons prévu trois modes de recherche (figure 3), qui correspondent à autant de compromis entre puissance et simplicité.

- **Recherche sémantique** : c'est le mode par défaut, le plus simple mais aussi le plus limité. L'utilisateur n'aura qu'à choisir une recherche prédéfinie dans une liste.
- **Recherche libre** : ce mode permet, à l'aide d'un assistant, de composer progressivement une requête plus ou moins complexe, correspondant à la plupart des besoins constatés chez les utilisateurs. Privilégiant la simplicité, elle ne donne pas accès à toutes les fonctionnalités du moteur de recherche.
- **Recherche avancée** : ce mode permet d'utiliser un langage de requête de manière « classique ».

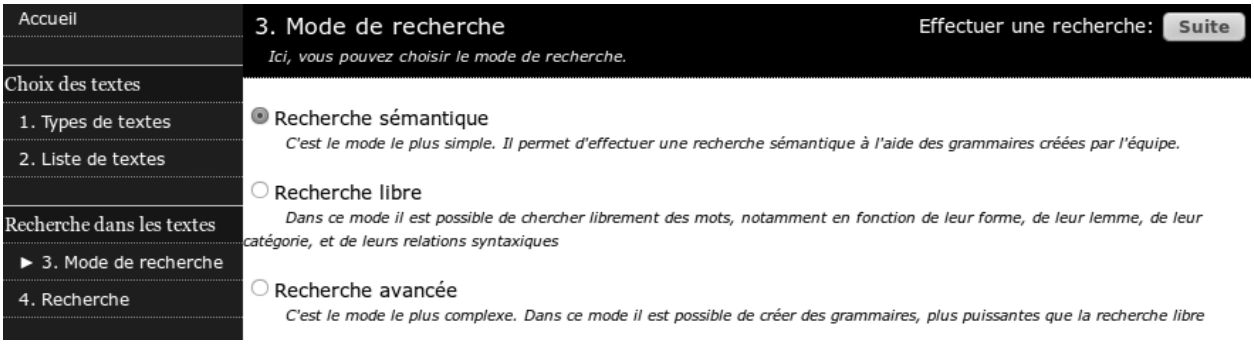

Figure 3: choix du mode de recherche.

Les deux premiers modes de recherche sont compatibles avec le mode de recherche avancée : il est possible de débuter une requête dans l'un de ces deux premiers modes, puis de l'étendre en mode avancé. Cette fonctionnalité est particulièrement utile pour permettre aux utilisateurs de maîtriser le langage de requête, et de devenir ainsi progressivement des utilisateurs experts.

Nous allons illustrer cette étape à travers quelques exemples issus du présent ouvrage.

4.3.1 Étude de la citation dans l'écrit scientifique : recherche sémantique et concordancier

Dans son étude de l'utilisation de citations dans les prises de position, Magda Florez utilise le mode sémantique de ScienQuest. Après avoir défini un sous corpus, elle utilise une grammaire prédéfinie afin d'extraire toutes les citations du corpus, qui lui serviront de matériau de base pour la validation de ses hypothèses.

La recherche sémantique (figure 4) permet d'accéder à des occurrences en corpus, à partir d'une vingtaine de grammaires prédéfinies, élaborées par l'équipe en utilisant le mode avancé. Les grammaires sont groupées par thématiques : la dénomination d'entités dans les textes, la formulation d'hypothèses, les prises de position de l'auteur, les propositions propres de l'auteur et les citations. C'est une grammaire de cette dernière catégorie qu'utilise Magda Florez, afin de lister les occurrences de citations de son sous-corpus.

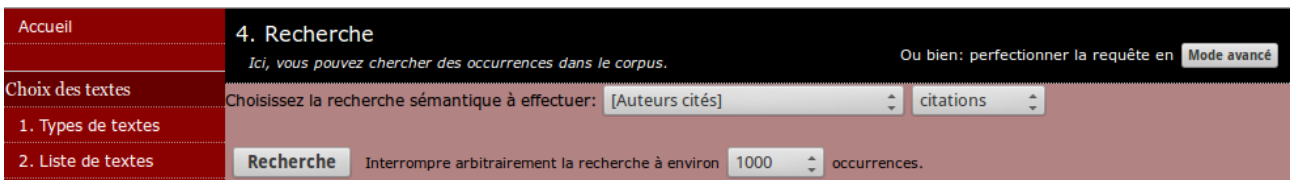

Figure 4: recherche sémantique d'occurrences de citations.

Chaque thème de recherche est adossé à une grammaire (invisible pour l'utilisateur), qui va effectuer des recherches avancées de manière transparente pour l'utilisateur.

4.3.2 Étude des verbes de constat : recherche libre

Pour les utilisateurs souhaitant aller plus loin que les grammaires prédéfinies que propose le mode sémantique. la recherche libre permet des formes, lemmes et/ou catégories, ainsi que des relations syntaxiques. Suivant en cela notre approche progressive, l'interface se présente au départ sous la forme d'un simple champ textuel (voir figure 5), permettant de rechercher un seul mot en fonction de sa forme textuelle. Des boutons et menus déroulants permettent :

- de choisir d'autres critères pour ce mot (lemme, partie du discours, flexion) ;
- d'ajouter d'autres mots.

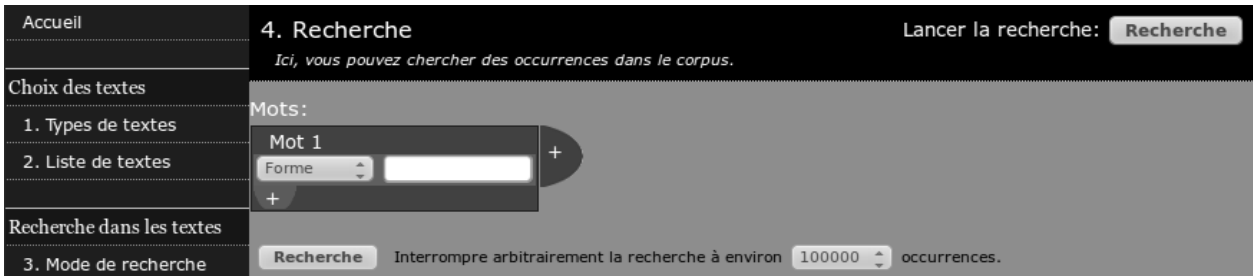

Figure 5: recherche libre : situation initiale.

Lorsqu'au moins deux mots sont utilisés, un champ apparaît, permettant facultativement de préciser les relations syntaxiques entre les mots. Ainsi, à partir d'une base simple constituée d'un seul champ, l'utilisateur peut étendre le formulaire pour intégrer de nouvelles contraintes.

Dans son étude des verbes de constat dans les écrits scientifiques, Francis Grossman étudie la distribution du verbe *voir* avec comme sujet le pronom *nous* ou *on*. Ces occurrences peuvent être recherchées en effectuant deux requêtes distinctes, une avec *nous* (figure 6) et une avec *on*. On peut aussi, au prix d'un peu de complexité, utiliser une expression régulière pour effectuer une seule requête (figure 7). La recherche n'est pas limitée à deux mots. Ainsi, il est possible d'extraire les occurrences de *pouvoir voir* avec comme sujet le pronom *nous* ou *on*. Mais autant la relation entre *nous* et *voir* est simple (relation sujet) à déterminer, autant la relation entre *pouvoir* et *voir* est difficile à nommer, si l'on ne connaît pas le modèle utilisé par l'analyseur syntaxique (en l'espèce, il s'agit simplement d'une relation objet). Or, l'utilisateur n'est pas supposé connaître ce modèle. Pour ce cas de figure, ScienQuest dispose d'une relation syntaxique générique, « n'importe quelle relation », qui épargne à l'utilisateur d'avoir à se plonger dans le modèle syntaxique de l'analyseur (figure 8).

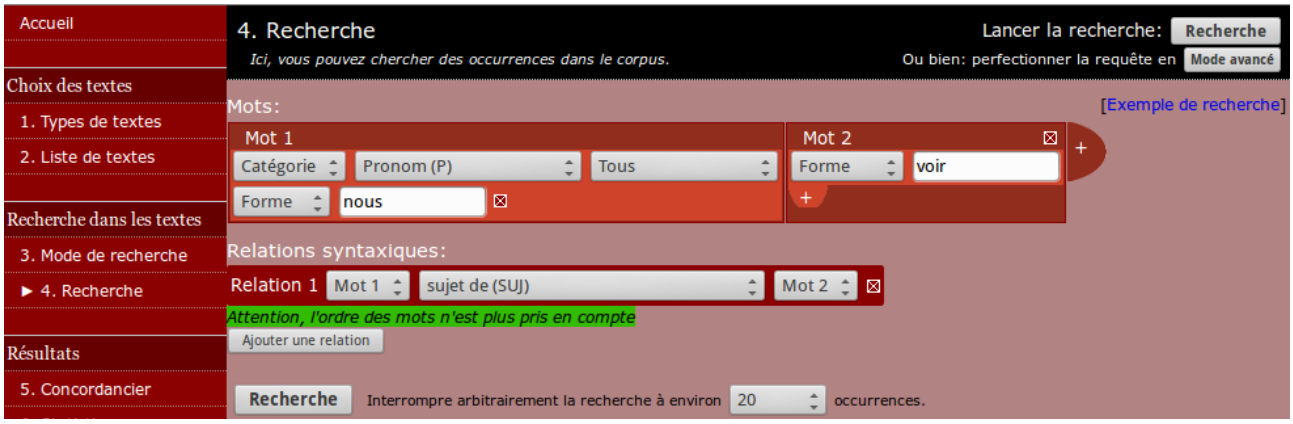

Figure 6: recherche de voir *avec comme sujet le pronom* nous.

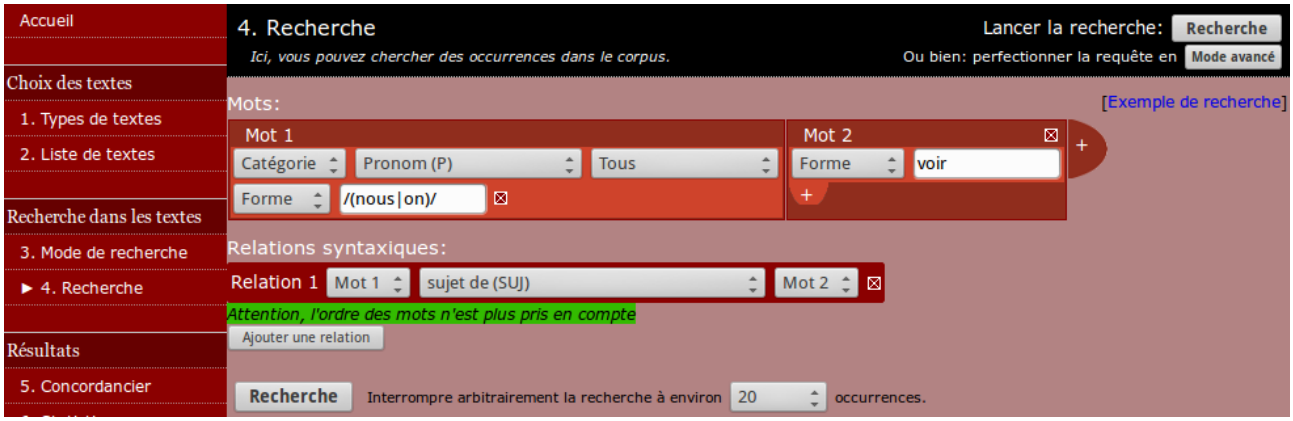

Figure 7: recherche de voir *avec comme sujet le pronom* nous *ou* on.

Figure 8: recherche de pouvoir voir *avec comme sujet le pronom* nous *ou* on.

Durant les tests de la plateforme avec des utilisateurs, nous avons constaté que la question de la linéarité de la requête, c'est à dire si les mots de la requête se suivent ou non, restait souvent assez floue pour les utilisateurs. Avec des corpus non arborés, il est d'usage d'effectuer des requêtes linéaires : les mots de la requête doivent se suivre, ou bien être séparés par moins d'un certain nombre de mots « blancs », qui seront ignorés. Quoi qu'il en soit, l'ordre des mots sera respecté : les premier mot de la requête sera toujours le premier mot dans les occurrences trouvées ; de même pour le deuxième, le troisième, etc. Lors de nos tests avec des utilisateurs, cette approche s'est révélée assez intuitive. Dans un corpus arboré, on ajoute une possibilité : deux mots peuvent être reliés par une relation syntaxique. Dans ce cas, il n'est pas nécessaire d'inclure de mots « blancs », et l'ordre des mots n'est pas nécessairement respecté. Ces possibilités peuvent se combiner de nombreuses manières ; on peut par exemple avoir une portion de requête dans laquelle les mots se suivent, et dont l'un de ces mots est relié à un autre par une relation syntaxique. Nous avons remarqué que cette notion était difficile à maîtriser pour des utilisateurs non-experts. Nous avons dans un premier temps choisi de nous limiter à un choix de linéarité global : soit tous les mots de la requête se suivent, soit on se base uniquement sur les relations syntaxiques ; mais cela s'est avéré encore trop complexe pour de nombreux utilisateurs. Nous avons donc décidé de réduire encore les possibilités, de la manière la plus intuitive possible pour les utilisateurs :

- lorsqu'aucune relation syntaxique n'est utilisée, les mots se suivent comme dans un corpus non-arboré ;
- lorsqu'au moins une relation syntaxique est utilisée, on se base uniquement sur les relations syntaxiques, et les mots ne se suivent pas nécessairement.

4.3.3 Recherche avancée

Les modes précédents font le choix de la simplicité, au prix d'une certaine simplification. Le langage de requête reste néanmoins indispensable pour des besoins spécifiques ou des recherches approfondies, pour les utilisateurs avancés. Le mode de recherche avancée (figure 10) est conçu pour ces utilisateurs. Il permet de créer des requêtes dans le langage ConcQuest (Kraif 2008), l'un

des rares langages de requête prenant en charge les relations syntaxiques, langage que nous avons étendu pour permettre de créer de petites grammaires.

Ce mode permet de préciser de manière fine la linéarité entre mots, d'utiliser des variables, des listes, des opérateurs booléens, des quantifieurs, et de redéfinir les relations syntaxiques. Cette dernière possibilité est très intéressante pour s'affranchir des limitations de l'analyse syntaxique de surface fournie par les analyseurs syntaxiques actuels. Elle est utilisée dans de la figure 11, où l'on définit une nouvelle relation OBJPASSIF (objet passif) en combinant les relations SUJ (sujet) et AUX (auxiliaire).

```
(OBJPASSIF,#3,#1) = (SUJ,#1,#2)(AUX,#3,#2)
$formul=avancer,émettre,effectuer,faire,formuler,poser,prendre,proposer,retenir,soutenir,ut
iliser
Main = <lemma=$formul,#1> && <lemma=hypothèse,#2>::(OBJ,#1,#2) OR (OBJPASSIF,#1,#2) OR
(ADJ,#2,#1)
```

Figure 9: exemple de grammaire.

Les grammaires utilisées dans le mode sémantique sont conçues avec ce langage.

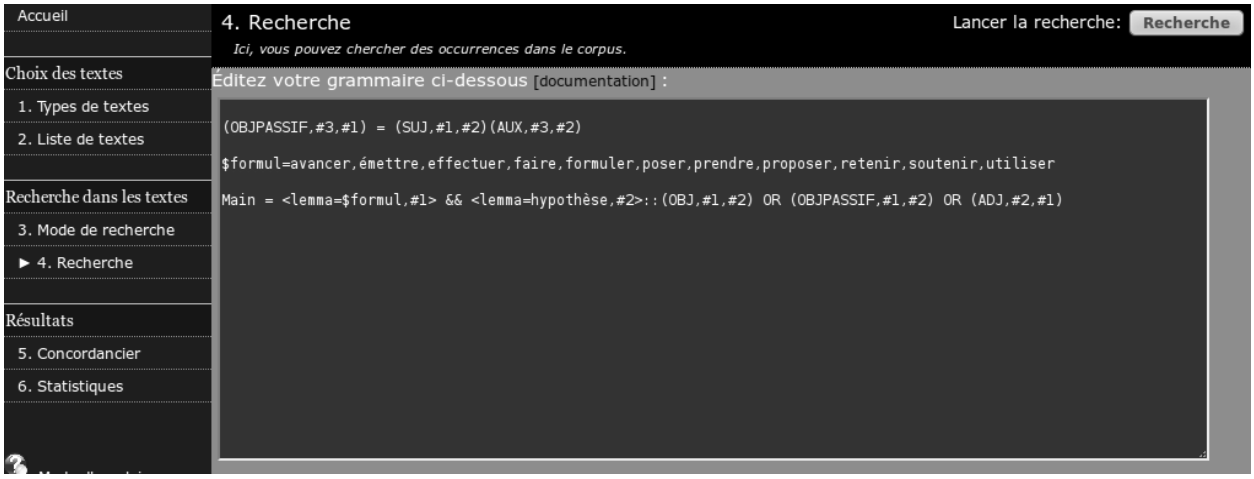

Figure 10: recherche avancée.

4.4 Consultation des résultats

Les résultats d'une recherche sont exploitables de trois manières.

La vue par défaut est une vue en concordancier, dans laquelle les occurrences trouvées s'affichent avec leur contexte. Cette vue est utile pour une étude qualitative.

Une autre vue permet de consulter des statistiques sur les occurrences trouvées, dans un cadre plus quantitatif. Elle permet de consulter les fréquences d'apparition des différentes occurrences, de manière globale dans tout le sous-corpus, ou bien de manière contrastive entre les différents types de textes et de parties textuelles.

Enfin, il est possible d'exporter les résultats, soit vers un tableur (format CSV), soit sous forme de tableau HTML (exploitable directement, ou bien importable dans d'autres logiciels).

4.4.1 Vue en concordancier

Une première vue des résultats est affichée directement sous le formulaire de recherche. La figure 11 montre un exemple de vue en concordancier dans le cadre de l'étude des citations positionnelles par Magda Florez. Le court contexte affiché autour des occurrences trouvées ne suffit pas toujours, notamment pour bien appréhender l'environnement rhétorique d'une occurrence, qui se déploie souvent sur plusieurs phrases, c'est pourquoi il est possible d'élargir ce contexte en cliquant dessus : on voit alors les 200 mots précédant et suivant l'occurrence (figure 12). La mise en forme du texte original est préservée dans une certaine mesure (paragraphes, italique, etc.) afin de faciliter

la lecture.

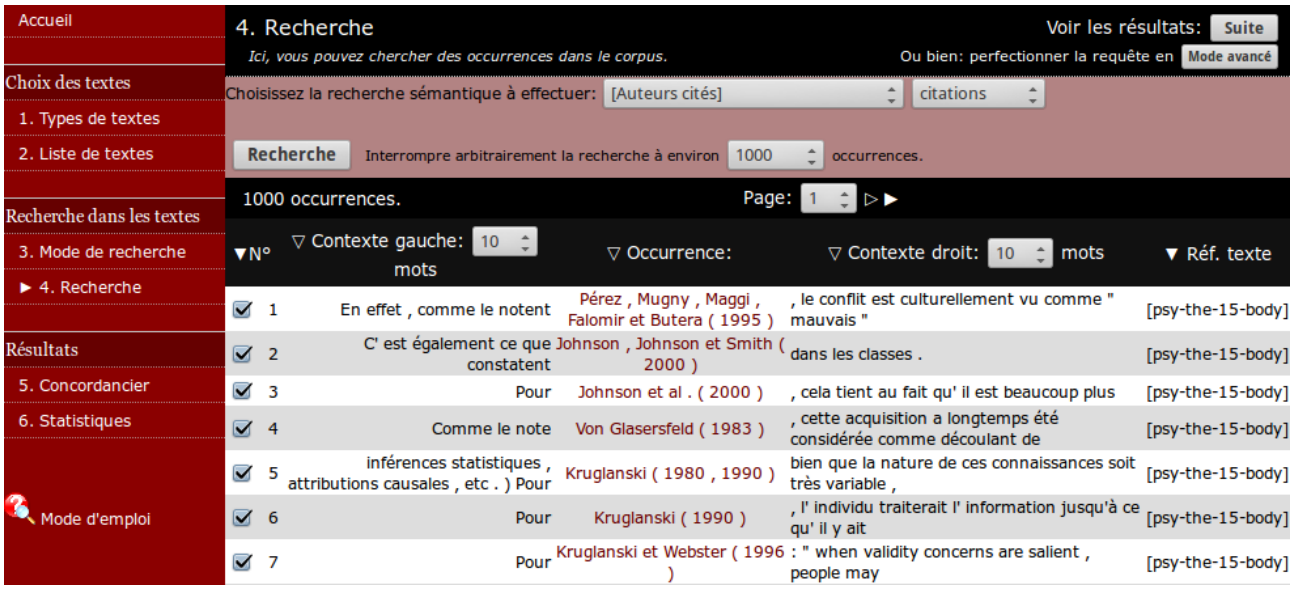

Figure 11: visualisation en concordancier des occurrences de citations.

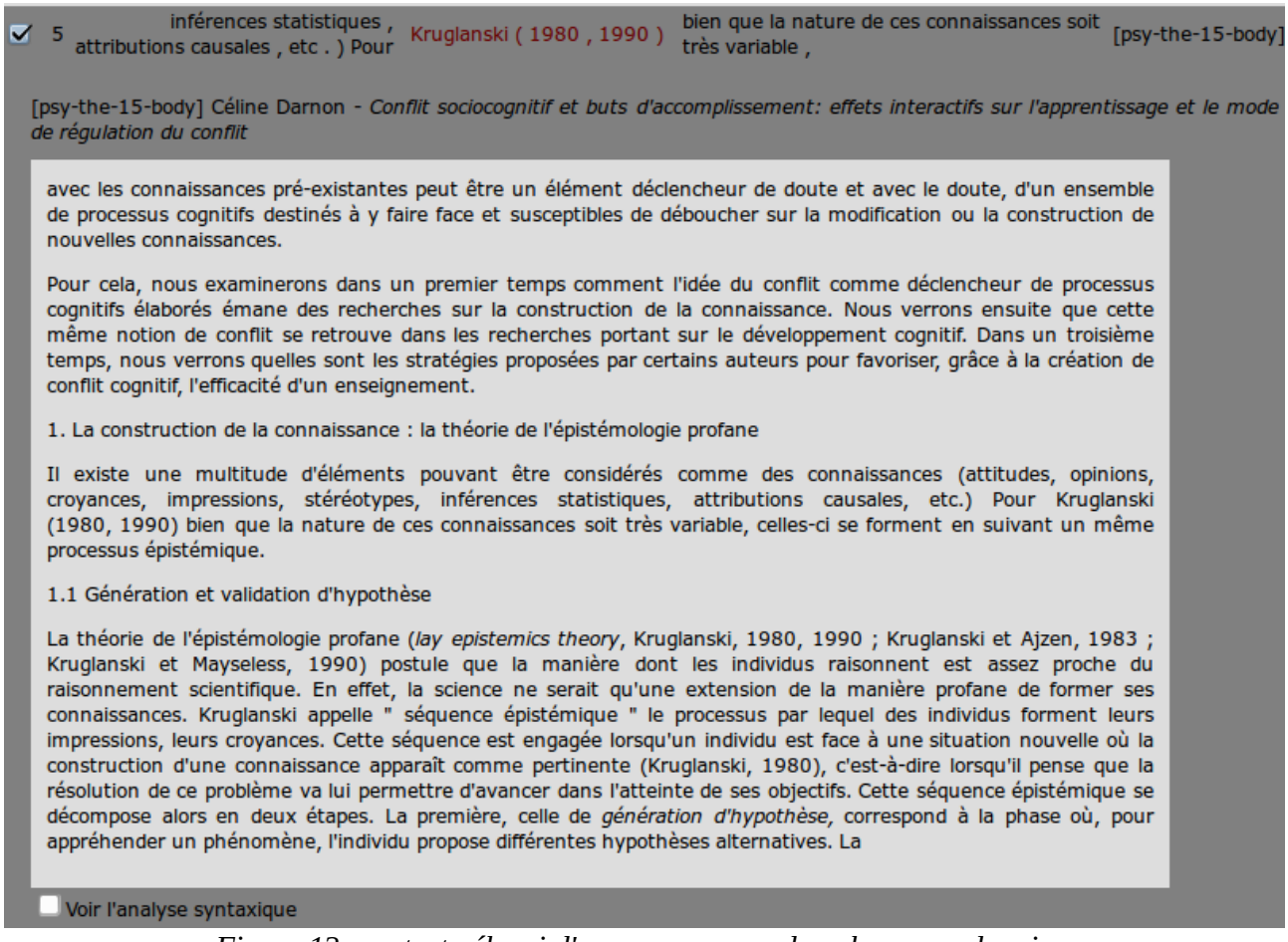

Figure 12: contexte élargi d'une occurrence dans le concordancier.

Grâce au concordancier et au contexte élargi, l'utilisateur peut filtrer les résultats (à l'aide d'une case décochable dans chaque ligne du concordancier), afin de ne conserver que ceux qui sont pertinents pour son étude. Ainsi, dans le cas de Magda Florez, la grammaire qu'elle a utilisée retourne toutes les citations, indépendamment du fait qu'elles fassent partie d'une prise de position. En décochant les cases correspondant aux occurrences ne faisant pas partie d'une prise de position, elle a ainsi pu conserver uniquement les citations pertinentes pour son travail. Après cette tâche d'extraction de données facilité par ScienQuest, elle a ensuite pu exporter les résultats dans un tableur (les formats CSV et HTML sont supportés), afin de passer au travail proprement linguistique

d'étude des schémas syntaxiques.

De plus, comme nous l'avons déjà signalé, les annotations du corpus, qui sont obtenues automatiquement, sont sujettes à des erreurs. Ces erreurs sont facilement repérables par un utilisateur humain, mais si elles sont nombreuses, elles peuvent totalement fausser les résultats. Par exemple, dans la phrase « *on pourrait même dire que l'ensemble des lexies EST la langue* », la forme *EST* est analysée par erreur comme un nom (le point cardinal), probablement parce qu'il apparaît en majuscule dans le texte. Ou encore, dans la phrase « *les élèves doivent construire la RGC de la fonction produit f(g)* », la forme *produit* est analysée par erreur comme un verbe conjugué. C'est pourquoi l'utilisateur a la possibilité de neutraliser les occurrences qui posent problème, simplement en décochant la ligne de cette occurrence dans l'interface. L'occurrence en question sera alors décomptée des résultats (concordances comme statistiques).

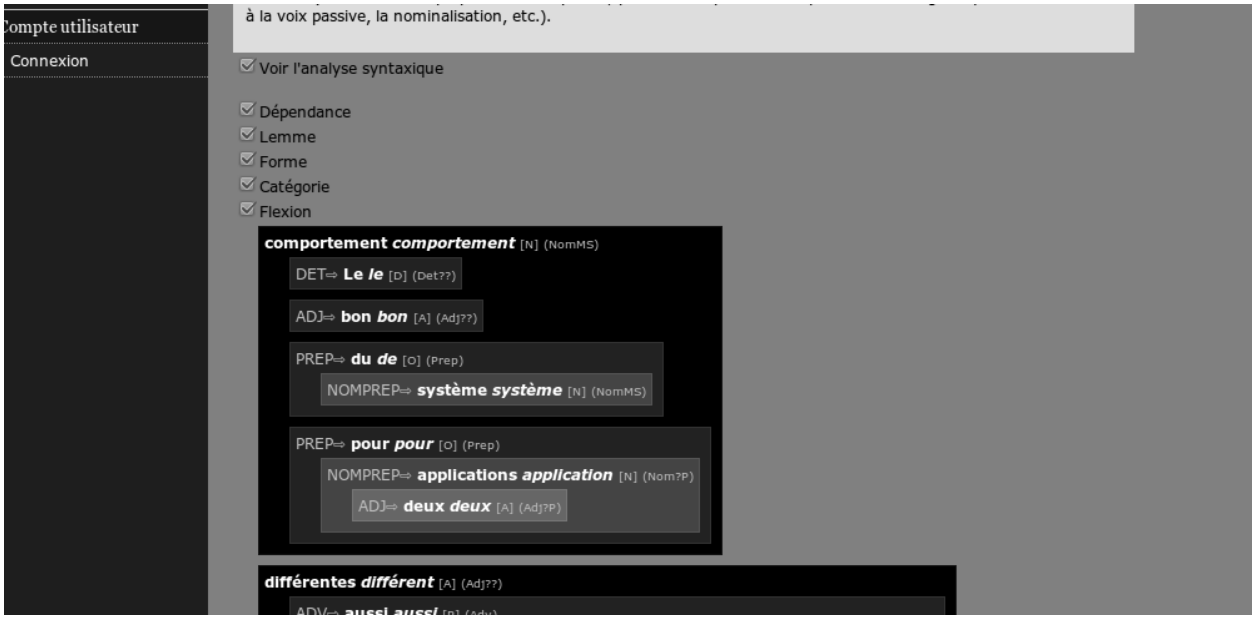

Figure 13: visualisation KWIC de l'analyse syntaxique.

Pour les utilisateurs avancés, il est possible de consulter les annotations du corpus pour chaque phrase contenant une occurrence (figure 13). Cela permet d'étudier la manière dont le corpus est effectivement annoté, et ainsi d'affiner les recherches.

4.4.2 Vue statistique

Des statistiques sur les occurrences trouvées sont disponibles : le nombre d'occurrences et le pourcentage des lemmes et des formes, ainsi que leur distribution par discipline, genre textuel, partie textuelle.

Le pourcentage des lemmes et des formes est utile dans le cas de requêtes complexes. Si l'on prend l'exemple de l'étude de Francis Grossmann sur les verbes de constat, avec une requête telle que « *voir* avec *nous* ou *on* comme sujet », cela permet de voir immédiatement quelle est la proportion d'occurrences avec *on* par rapport à *nous* (figure 14).

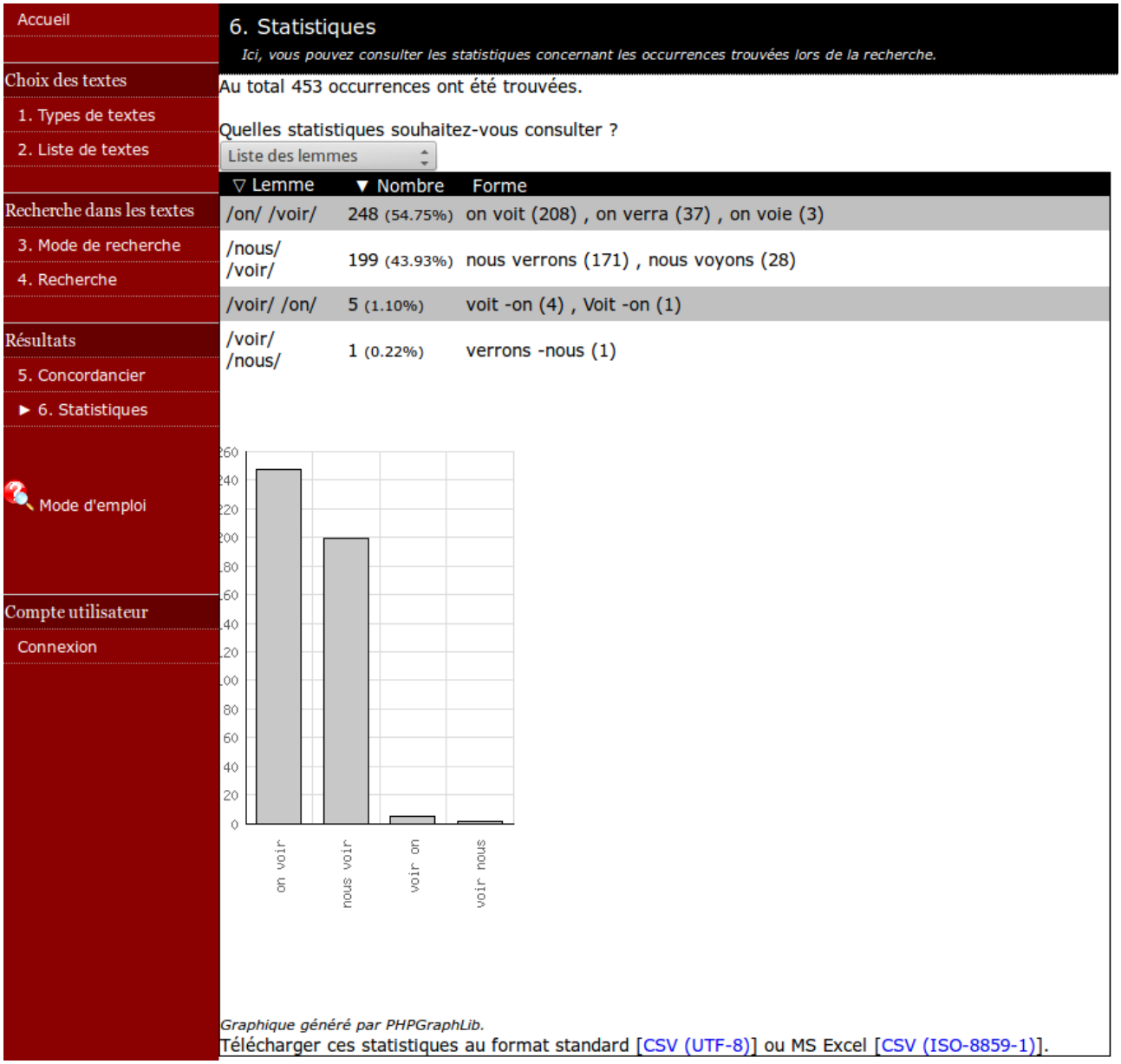

Figure 14: répartition en lemmes des résultats de « voir *avec* nous *ou* on *comme sujet ».*

Les informations sur la distribution sont évidemment précieuses pour l'étude distributionnelle. Par exemple, elles permettent de constater que les occurrences de « *voir* avec *nous* ou *on* comme sujet » sont proportionnellement plus rares dans les conclusions et les annexes que dans le reste des textes (figure 15). Autre exemple : Magda Florez a ainsi pu, après avoir filtré manuellement les citations extraites par ScienQuest pour ne retenir que celles qui faisaient partie d'une prise de position, déterminer dans quelles disciplines, types de textes et parties textuelles elles se situaient.

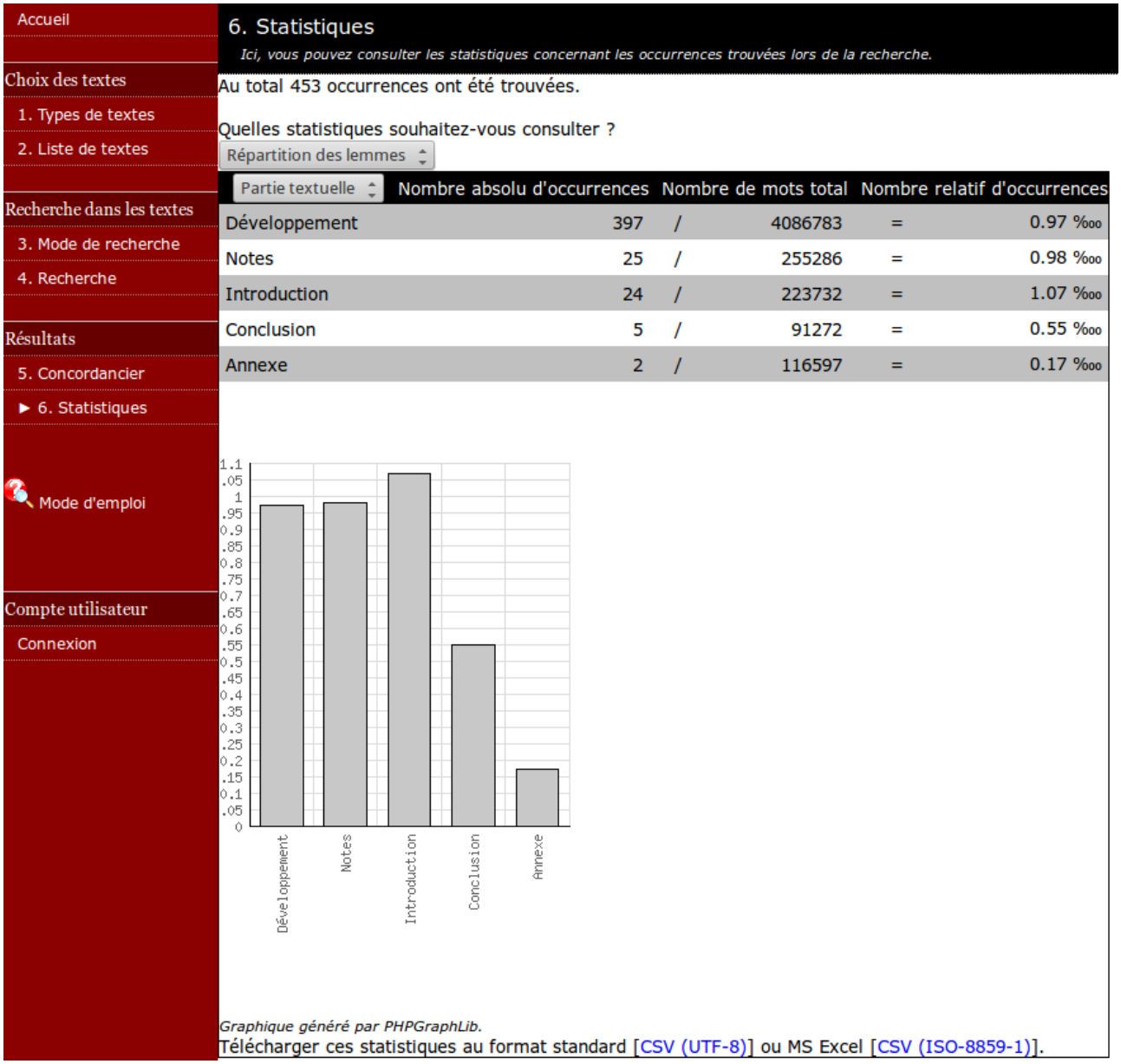

*Figure 15: répartition en parties textuelles des résultats de « voir *avec* nous *ou* on *comme sujet* ».*

5 Bilan et conclusion

Sur la période allant du début du lancement public du site fin juin 2010 à fin juillet 2012, 9021 requêtes ont été effectuées (en 1474 sessions). Le mode guidé est utilisé pour 75% des requêtes, le mode sémantique (grammaires locales prédéfinies) pour 23% et le mode avancé pour 2% ; cela démontre bien selon nous l'intérêt de ces deux premiers modes de recherche. Malgré tout, il reste que les connaissances d'ordre syntaxique présentent une complexité inhérente qui freine quelque peu l'utilisation grand public de tels corpus, puisque seulement 47% des requêtes guidées comportaient des contraintes d'ordre syntaxique.

L'utilisation du système Scientext dépasse aujourd'hui le cadre du projet ANR dont il est issu. Il est par exemple utilisé en didactique du FLE dans le cadre du projet FULS[8], et intègre de nouveaux corpus, traités avec un analyseur différent, pour le projet ANR Emolex[9].

Plusieurs améliorations du système sont prévues : l'ajout de nouveaux corpus, pour le mode guidé l'ajout de nouvelles fonctionnalités (par exemple une présélection des relations syntaxiques à partir des parties du discours choisies). Ces améliorations seront testées sur un ensemble

8 http://scientext.msh-alpes.fr/fuls/

9 http://www.emolex.eu/

d'utilisateurs non informaticiens.

Bibliographie

ABEILLÉ A., CLÉMENT L., TOUSSENEL F. (2003). « Building a treebank for French ». ABEILLÉ A. (ed) *Treebanks*. Dordrecht : Kluwer.

BICK Eckhard (2004). « Parsing and evaluating the French Europarl corpus », PAROUBEK Patrick, ROBBA Isabelle & VILNAT Anne (ed.): *Méthodes et outils pour l'évaluation des analyseurs syntaxiques* (Journée ATALA, 15 mai 2004). p. 4-9. Paris: ATALA.

BICK Eckhard (2005). « Live use of Corpus data and Corpus annotation tools in CALL : Some new developments in VISL ». HOLMBOE Henrik (ed.), *Nordic Language Technology, Årbog for Nordisk Sprogteknologisk Forskningsprogram* 2000-2004 (Yearbook 2004), p. 171-186. Copenhague, Danemark : Museum Tusculanum.

BOURIGAULT Didier (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Mémoire de HDR. Toulouse.

CHRIST Oli (1994). « A modular and flexible architecture for an integrated corpus query system », *Proceedings of COMPLEX'94*, Budapest.

CHRIST Oli, SCHULZE B.M. (1995). « Ein flexibles und modulares Anfragesystem für Textcorpora. Tagungsbericht des Arbeitstreffen Lexikon + Text ». Tübingen, Allemagne : Niemeyer.

KRAIF Olivier (2008), « Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest », *Actes des 9ème Journées d'analyse statistique des données textuelles*, JADT 2008, p. 625-634. Lyon: Presses universitaires de Lyon.

LEZIUS Wolfgang, KÖNIG Esther (2000). « Towards a search engine for syntactically annotated corpora ». SCHUKAT-TALAMAZZINI Ernst G, ZÜHLKE Werner (ed.): *KONVENS-2000 Sprachkommunikation*, p. 113-116. Ilmenau, Allemagne : VDE-Verlag.

LEZIUS Wolfgang (2002) *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora* (German) Ph.D. thesis, IMS, University of Stuttgart Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 8, number 4.

MERTENS Piet (2002). « Les corpus de français parlé ELICOP : consultation et exploitation ». BINON Jean, DESMET Piet, ELEN Jan, MERTENS Piet, SERCU Lies (ed.) *Tableaux vivants*, Opstellen over taal- en onderwijs, aangeboden aan Mark Debrock, Symbolae, Facultatis Litterarum Lovaniensis, Series A, vol. 28. 383-415. Louvain, Belgique : Leuven Universitaire Pers.

SILBERZTEIN Max (2006). « NooJ's Linguistic Annotation Engine ». KOEVA S., MAUREL D., SILBERZTEIN M. (ed.), *INTEX/NooJ pour le Traitement Automatique des Langues*. Cahiers de la MSH Ledoux. Presses Universitaires de Franche-Comté, pp. 9-26.

TUTIN Agnès, GROSSMANN Francis, FALAISE Achille, KRAIF Olivier (2009). *Autour du projet Scientext : étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques*. Journées Linguistique de Corpus. 10-12 septembre 2009, Lorient.

VOORMANN Holger (2002) *TIGERin - Grafische Eingabe von Suchanfragen in TIGERSearch.* Diploma thesis. Fakultät Informatik, Universität Stuttgart.