ScienQuest: a treebank exploitation tool for non NLP-specialists

Achille Falaise[1], Olivier Kraif[2], Agnès Tutin[2], David Rouquet[1]
(1) LIG-GETALP, University of Grenoble, France
(2) LIDILEM, University of Grenoble, France
achille.falaise@imag.fr, olivier.kraif@u-grenoble3.fr,
agnes.tutin@u-grenoble3.fr, david.rouquet@imag.fr

ABSTRACT

The exploitation of syntactically analysed corpora (or treebanks) by non NLP-specialist is not a trivial problem. If the NLP community wants to make publicly available corpora with complex annotations, it is imperative to develop simple interfaces capable of handling advanced queries. In this paper, we present query methods developed during the Scientext project and intended for the general public. Queries can be made using forms, lemmas, parts of speech, and syntactic relations within specific textual divisions, such as title, abstract, introduction, conclusion, etc. Three query modes are described: an assisted query mode in which the user selects the elements of the query, a semantic mode which includes local pre-established grammars using syntactic functions, and an advanced search mode where the user create custom grammars.

ScienQuest: un outil d'exploitation de corpus arborés pour les non spécialistes du TALN

RÉSUMÉ

L'exploitation de corpus analysés syntaxiquement (ou corpus arborés) pour le public non spécialiste du TALN n'est pas un problème trivial. Si la communauté du TALN souhaite mettre à la disposition des chercheurs non-informaticiens des corpus comportant des annotations linguistiques complexes, elle doit impérativement développer des interfaces simples à manipuler mais permettant des recherches fines. Dans cette communication, nous présentons les modes de recherche « grand public » développés dans le cadre du projet Scientext, qui met à disposition un corpus d'écrits scientifiques interrogeable par section textuelle, par partie du discours et par fonction syntaxique. Trois modes de recherche sont décrits : un mode libre et guidé, où l'utilisateur sélectionne lui-même les éléments de la requête, un mode sémantique, qui comporte des grammaires locales préétablies à l'aide des fonctions syntaxiques, et un mode avancé, dans lequel l'utilisateur crée ses propres grammaires.

KEYWORDS : corpus exploitation environments, treebanks, assisted grammar creation, visualization of linguistic information.

MOTS-CLÉS : environnement d'étude de corpus, corpus étiquetés et arborés, création de grammaires assistée, visualisation d'information linguistique.

1 Version courte en français

1.1 Introduction

Les outils d'exploration de corpus annotés, en particulier de corpus arborés (c'est-à-dire comportant des relations syntaxiques), sont souvent complexes à utiliser, *a fortiori* pour des utilisateurs non initiés à la linguistique-informatique. L'ergonomie et la facilité d'utilisation des outils sont cependant des enjeux majeurs en TALN, surtout si l'on souhaite diffuser des traitements et des annotations linguistiques complexes dans la communauté des linguistes. Pour élargir le nombre d'utilisateurs des corpus annotés, il est essentiel de développer des outils d'exploration de corpus faciles à manipuler mais puissants. C'est ce qui nous a amenés à proposer un environnement de recherche simple, adapté aux linguistes, didacticiens, lexicographes ou épistémologues.

Nous présentons ici l'outil développé dans le cadre du projet Scientext[1], qui propose des modes de recherche simples pour non spécialistes du TALN sur un corpus d'écrits scientifiques analysé syntaxiquement. Il s'agit d'un outil d'étude en ligne de corpus arborés construit à partir d'un scénario de recherche simple : choix d'un corpus, recherche de phénomènes linguistiques, et enfin affichage des résultats. Ce scénario de base est facile à appréhender, et se décompose en plusieurs écrans simples qui peuvent s'enrichir de fonctions plus complexes « en douceur ».

Dans un premier temps, nous présentons les outils existants pour l'étude de corpus arborés, en particulier pour le français, rares et peu conviviaux. Nous détaillons ensuite les fonctionnalités de notre outil, et effectuons enfin un bilan de son utilisation.

1.2 Les corpus

Dans le cadre du projet Scientext, quatre corpus de textes scientifiques ont été collectés. Deux des corpus contiennent des textes anglais, et les deux autres des textes français. Ces corpus sont librement consultables sur le site du projet. D'autres corpus de textes littéraires et journalistiques en allemand, anglais, espagnol, français et russe ont été collectés dans le cadre du projet EMOLEX[2], mais ne sont pas consultables pour des raisons de droits. Tous ces corpus ont été annotés morpho-syntaxiquement, à l'aide de divers analyseurs (Syntex, Connexor, XIP, DeSR), et ont pu être exploités à l'aide de ScienQuest.

1.3 Les modes d'accès aux textes

Après avoir sélectionné un sous-corpus en fonction des genres et des sections textuelles désirés (figure 2), l'utilisateur peut effectuer des recherches sur le corpus selon trois modes, de complexité et d'expressivité croissante :

- **Un mode sémantique** permet d'accéder à des occurrences en corpus, à partir de grammaires prédéfinies. Les grammaires sont définies à l'aide d'un langage de requête existant (ConcQuest — Kraif, 2008), que nous avons étendu.

[1] http://scientext.msh-alpes.fr
[2] http://www.emolex.eu

- **Un mode simple et guidé** avec un assistant permet à l'utilisateur de sélectionner des formes, lemmes et/ou catégories, ainsi que les relations syntaxiques désirées (figures 4 et 5).
- **Un mode complexe** permet d'accéder à des occurrences en corpus, à partir de grammaires, utilisant les dépendances syntaxiques, les relations linéaires et des variables.

Une fois la requête effectuée, les occurrences trouvées sont affichées soit dans un concordancier (figure 6), soit sous forme de statistiques distributionnelles.

1.4 Conclusion

L'utilisation du système Scientext dépasse aujourd'hui le cadre du projet dont il est issu. Il est par exemple utilisé en didactique du FLE dans le cadre du projet FULS[3], et intègre de nouveaux corpus, traités avec des analyseurs différents, pour le projet EMOLEX.

Depuis le lancement public du site fin juin 2010, 9 021 requêtes ont été effectuées (en 1 474 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.

Plusieurs améliorations du système sont prévues : l'ajout de nouveaux corpus et l'ajout de fonctionnalités pour le mode guidé. Ces améliorations seront testées sur un ensemble d'utilisateurs non spécialistes du TALN.

[3] http://scientext.msh-alpes.fr/fuls

2 Introduction

Textual corpora are more and more often enriched with different types of linguistic annotations, such as structural and discursive annotations, lemmatisation, part-of-speech (POS) tagging, or syntactic tree structures. These annotations may be very appealing for non-NLP specialists, such as linguists, language teachers, lexicographers, and epistemologists. However, exploration tools for such corpora, especially treebanks (i.e. with syntactic relations) are often complex to use, a fortiori for users not familiar with computational linguistics. In order to broaden the scope of access to these annotations outside the NLP community, it is essential to develop tools that are both powerful, but easy to handle for corpus exploration. This objective led us to propose ScienQuest, a simple research environment suitable for non NLP-specialists.

ScienQuest was developed in the context of the Scientext project[4]. It is an online generic tool, which focuses on a GUI suitable for non NLP-specialists. It is based on the ConcQuest (Kraif, 2008) command-line search engine. ScienQuest was first used as part of the Scientext project, which includes several corpora of scientific texts, and then for several new corpora. ScienQuest is based on a simple search scenario. After building a sub-corpus based on the metadata of the texts in the corpus, user requests within this sub-corpus provide results displayed in two fashions: Key Word In Context (KWIC) and distributional statistics. This straightforward three-part baseline scenario is represented within the interface by the division into several simple screens.

In this paper, we first present the corpora currently operated within ScienQuest. Then, we describe briefly some of the existing tools for the study of treebanks. Finally, we detail the features of ScienQuest, and conclude with a review of its use.

3 Of corpora and users

The corpora of the Scientext project were collected to conduct a linguistic study on reasoning and positioning of authors, through phraseology, enunciative and syntactic markers related to syntactic causality. The corpora were parsed with Syntex (Bourigault, 2007). They consist of an English corpus of biology and medical articles (14M words), a corpus of argumentative texts of French learners of English (1M words), a French corpus of varied scientific texts (a range of genres and disciplines totalling 5M words), and a corpus of French scientific communication reviews (502 reviews, 35k words). These corpora are freely available within ScienQuest on the Scientext project website. A study is underway to draw upon these corpora using ScienQuest in the context of language courses.

ScienQuest has recently been integrated for the exploitation of the corpora of the EMOLEX[5] project, which aims to investigate the multilingual lexicon of emotions within five corpora of fiction and newspaper articles in English (200M words, analysed with XIP), French (230M words, analysed with Connexor), German (301M words, analysed with Connexor), Spanish (286M words, analysed with Connexor), and Russian (554k

[4] http://scientext.msh-alpes.fr
[5] http://www.emolex.eu

word, analysed with DeSR). For copyright issues, these corpora are unfortunately not publicly available yet.

ScienQuest was designed with usability in mind, and therefore was designed with user input in mind. A first survey was conducted in March 2008 with a group of researchers and students in linguistics, teaching, and communication from four different laboratories. A first prototype was built based on this first survey. A second survey was then conducted, with both new participants and researchers involved in the previous survey.

4 A brief overview of annotated corpora exploration tools

There are several environments for the study of linguistically annotated corpora. These environments are based on query languages; they sometimes come with a graphical environment, usually limited to a graphical query editor. This type of graphical environment can improve the readability compared to a query language used alone, but still maintains the same conceptual complexity. These tools require a familiarity with computer science basics such as logical operators and regular expressions that are often poorly mastered by non-specialists, who therefore are often reluctant to use these tools.

Most corpus research tools do not integrate the syntactic level. The Corpus Query Processor (CQP – Christ, 1994), developed at the *Institut für Maschinelle Sprachverarbeitung* of Stuttgart, has become a standard element in the community of NLP. A graphical interface for CQP, CQPWeb[6], is available. This GUI, has an interesting *simple mode*, where the user can type a sequence of word forms or lemmas ; however, it does not provide a simple way for more advanced searches involving POS or morphological features, for users who do not know the CQP language. In contrast, the rather less known GUI employed for the lemmatized and POS-tagged corpus Elicop (Mertens, 2002) was a source of inspiration for our interface. It is based on an easy-to-complete form and does not require prior knowledge of a query language (see Figure 1). The system does not, however, permit one to build a subcorpora and is also limited to four words without syntactic relations.

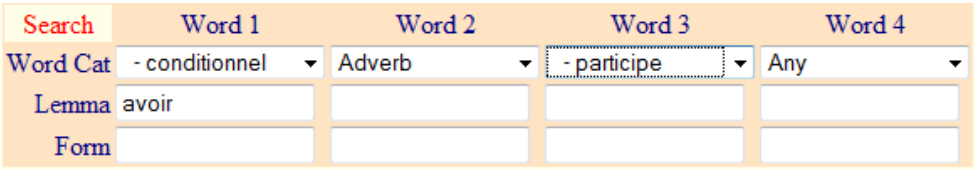

FIGURE 1 – Elicop search GUI.

TIGERSearch (Lezius and König, 2000) is one of the few graphic (off-line) environments for querying treebanks (of syntagmatic structure), but the tool is no longer maintained. This GUI is mostly a query editor, which makes querying more readable (especially for complex queries), but is not easier to master for users unfamiliar with computer science and NLP.

In conclusion, we can only deplore the lack of user-friendly online environments, especially those that incorporate treebanks. This gap is one of the reasons that led to the creation of ScienQuest.

[6] http://cwb.sourceforge.net/

5 ScienQuest features

Using Scientext consists of three main steps: sub-corpus construction, research, and finally the display of results. Unless otherwise stated, the examples given below are based on the corpus of English scientific texts.

5.1 Sub-corpus selection

By default, ScienQuest exploits the entire corpus. The first step is to simply accept this default choice or to select a sub-corpus. It is possible to group texts according to various criteria (e.g. text type, discipline, and textual section). For corpora with structural annotations (e.g. abstract, introduction, titles, etc.), it is also possible to restrict the sub-corpus according to these elements.

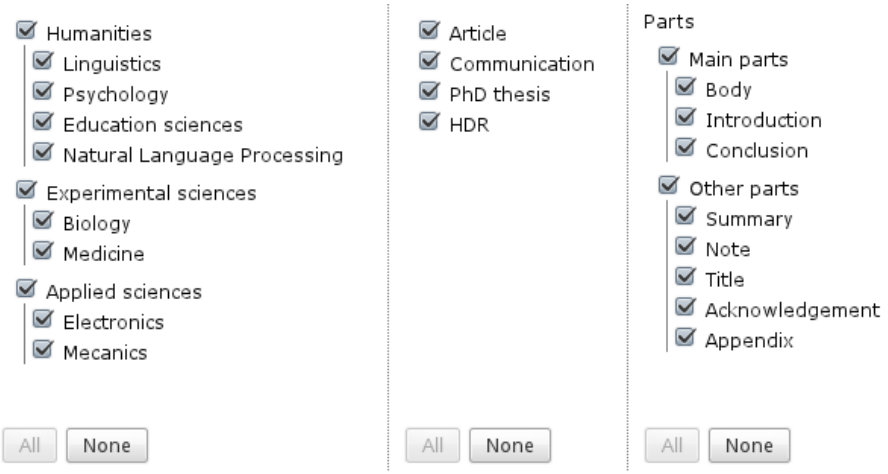

FIGURE 2 – Subcorpus selection GUI in ScienQuest for the French scientific texts corpus.

5.2 Search

Once the corpus is defined, the user is prompted to choose between three research modes. Whichever method chosen, the result of the user interaction will be a local grammar (local grammars are presented later in this paper). This grammar is compiled into the local query language used by the search engine, ConcQuest (Kraif, 2008) which performs searches within the corpus (see Figure 3).

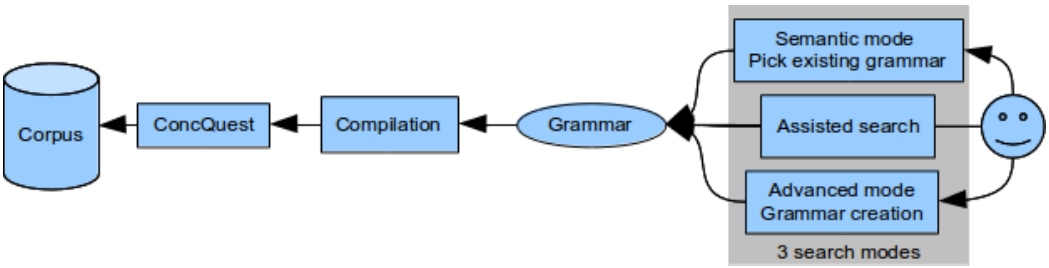

FIGURE 3 – Internal search process.

5.2.1 Semantic search: using a predetermined local grammar

A set of local grammars has been developed to enable semantic search within texts, so that the user is not encumbered by the complexity of queries. Fifteen local grammars were developed by (Tutin et al., 2009), primarily around the theme of reasoning and positioning of authors. The development of more local grammars is planned concerning other themes, especially to look for stereotypical expressions in a perspective of language teaching.

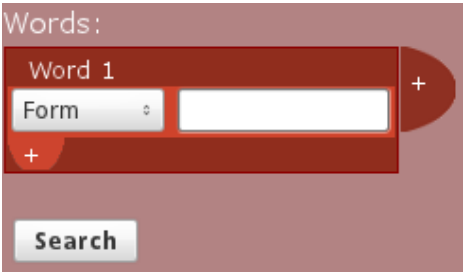

FIGURE 4 – Initial assisted search form.

5.2.2 Assisted search : an assistant for free searches

In assisted search mode, the interface first contains minimal content, with a single input field (search for one word, see Figure 4). Buttons allow the user to add words and constraints on form, lemma, part of speech (and possibly sub-category, e.g. proper noun, tensed verb, etc.). For more advanced users, regular expressions are accepted. By default, there are no syntactic relations between words, and their linear order is taken into account.

When several words are present, it is possible to specify a syntactic relation between these words. If a relation is selected, the word order is no longer taken into account (see Figure 5).

The form data are automatically converted into a local grammar in the back office. This mode is designed to satisfy most of the needs of medium-skilled users; however, it is deliberately limited to only a subset of features that can be easily intuitively understood, without the need of consulting the user guide. To use the full expressiveness of the search tool, one must switch to the advanced search option.

FIGURE 5 – Search for nouns which are direct object of the verb *to have.*

5.2.3 Advanced search: a local grammar language for treebanks querying

In the advanced search mode, the user can directly create a local grammar. This mode is dedicated to specialists, therefore we present only the main features here. The local grammar language is used to specify constraints on the words (form, lemma, part of speech, flexion), order, and syntactic relations between words. It is also possible to specify a list of words and variables.

Some innovative features of this language are specific to the treatment of treebanks, in particular the possibility to extend the syntactic relations encoded in the corpus. For example, the syntactic analyser Syntex, used for the Scientext corpora performs a shallow analysis, and thus creates no direct dependency relation between a verb in a perfect tense and its subject, but instead creates a relation SUBJ (subject) between the subject and auxiliary, and a relationship AUX (auxiliary) between the auxiliary and the verb. With ScienQuest local grammar language, it is possible to define a new generic relation or "deep subject" that takes into account this construction, as in the example grammar below. This grammar presents a set of rules detecting opinion verbs and their subject in deep syntax (e.g. the syntactic relation between [*I* or *we*] and [*to think*] in the sentence *we have thought* or *we can think*).

(SUBJINF,#2,#1) = (SUJ,#3,#1)(OBJ,#3,#2) *// For infinitive syntactic structures*
(SUBJAUX,#2,#1) = (SUJ,#3,#1) (AUX,#3,#2) *// For syntactic structures with auxilliaries*
(SUBJGENERIC,#2,#1) = (SUBJINF,#2,#1) OR (SUBJ,#1,#2) OR (SUBJAUX,#2,#1) *// New syntactic rel.*
$pron_author = we, I *// Pronouns refering to the author*
$v_opinion = adhere, admit, adopt, affirm, forward, hold, contradict, agree, criticize, suggest, defend, denounce, doubt, expect, estimate, judge, justify, think, postulate, prefer, favor, recognize, reject, refute, regret, reject, wish, stress, subscribe, support, suggest *// Opinion verbs*
Main = <lemma=$v_opinion,#1> && <lemma=$pron_author,#2> :: (SUBJGENERIC,#1,#2)

In order to help users to switch from semantic and assisted mode to the more complex advanced mode, all advanced queries can be initiated in semantic and assisted mode, and then enriched in advanced mode.

5.3 Visualisation of results

The search results are then browsable in a KWIC display (Figure 6), which includes information about the type of text, textual section, etc. In this view, the user can also remove the incorrect results, which will not appear in exports and statistics. This is very useful for fine-tuning results and for annotation error removal. These results can be exported in CSV and XLS formats as well as in HTML tables. It is also possible to broaden the context for a given line and to consult the syntactic dependencies.

Statistics on the occurrences found are available in both tabular and chart display: number of occurrences, percentage of lemmas and forms and their distribution by discipline, and text type or textual section. This type of functionality is still rarely available in tools dedicated to corpus study and is particularly interesting for the study of rhetorical structures in scientific writing.

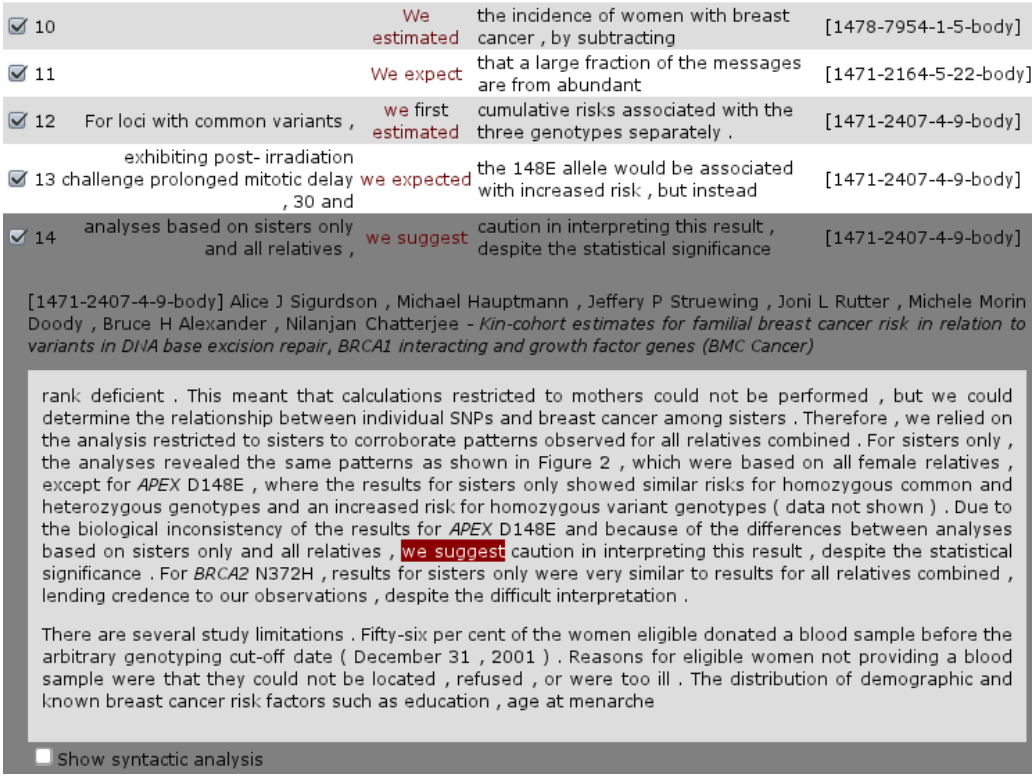

FIGURE 6 : KWIC visualisation of results.

6 First results and conclusion

Since the beginning of the public launch of ScienQuest in late June 2010, 9,021 requests were made during 1,474 sessions. Assisted search is used for 75% of the queries, semantic mode (predefined local grammars) for 23%, and advanced mode for the remaining 2%. We believe this demonstrates the importance of these first two search modes, which were introduced in ScienQuest. Nevertheless, the fact remains that syntactic knowledge has inherent complexity which hampers to a certain extent the use of treebanks, since only 47% of the assisted searches contained guided syntactic constraints.

The use of ScienQuest now exceeds the Scientext[7] project, from which it originated. It is for example used in the teaching of French as a foreign language in the FULS[8] project and incorporates new corpora for the EMOLEX[9] project.

Several system improvements are planned, such as adding new corpora and features for the assisted search. However, the ConcQuest search engine on which ScienQuest is based is rather slow ; we will envetually replace it with the new and faster ConcQuest 2.

[7] http://scientext.msh-alpes.fr
[8] http://scientext.msh-alpes.fr/fuls
[9] http://www.emolex.eu

References

Abeillé, A.; Clément, L.; Toussenel, F. (2003). Building a treebank for French. In Abeillé A. (ed) *Treebanks*. Dordrecht, Germany : Kluwer.

Bick, Eckhard (2004). Parsing and evaluating the French Europarl corpus. In Paroubek, Patrick; Robba, Isabelle and Vilnat, Anne (ed.): *Méthodes et outils pour l'évaluation des analyseurs syntaxiques* (Journée ATALA, 15 mai 2004). pp. 4-9. Paris, France: ATALA.

Bick, Eckhard (2005). Live use of Corpus data and Corpus annotation tools in CALL: Some new developments in VISL. In Holmboe, Henrik (ed.), *Nordic Language Technology, Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004* (Yearbook 2004), pp.171-186. Copenhagen, Danemark : Museum Tusculanum.

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

Christ, Oli (1994). A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94,* Budapest, Hungary.

Christ, Oli and Schulze, B.M. (1995). Ein flexibles und modulares Anfragesystem für Textcorpora. *Tagungsbericht des Arbeitstreffen Lexikon + Text*. Tübingen, Germany : Niemeyer.

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

Lezius, Wolfgang and König, Esther (2000). Towards a search engine for syntactically annotated corpora. In Schukat-Talamazzini, Ernst G. and Zühlke, Werner (ed.): *KONVENS-2000 Sprachkommunikation*, pp. 113-116. Ilmenau, Germany : VDE-Verlag.

Mertens, Piet (2002). Les corpus de français parlé ELICOP : consultation et exploitation. In 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, Belgium : Leuven Universitaire Pers.

Silberztein, Max. (2006). NooJ's Linguistic Annotation Engine. In Koeva, S.; Maurel, D. and 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 and Kraif, Olivier (2009). Autour du projet Scientext : étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques . In *Journées Linguistique de Corpus*. 10-12 septembre 2009, Lorient, France.