Intégration du corpus des actes de TALN à la plateforme ScienQuest

Achille Falaise[1]
(1) ICAR, ENS de Lyon, 15 parvis René Descartes, 69342 LYON cedex 7
achille.falaise@ens-lyon.fr

Résumé. Cette démonstration présente l'intégration du corpus arboré des Actes de TALN à la plateforme ScienQuest. Cette plateforme fut initialement créée pour l'étude du corpus de textes scientifiques Scientext. Cette intégration tient compte des méta-données propres au corpus TALN, et a été effectuée en s'efforçant de rapprocher les jeux d'étiquettes de ces deux corpus, et en convertissant pour le corpus TALN les requêtes prédéfinies conçues pour le corpus Scientext, de manière à permettre d'effectuer facilement des recherches similaires sur les deux corpus.

Abstract.

Integration of the TALN proceedings treebank to the ScienQuest platform
This demonstration shows the integration of the TALN proceedings Treebank to the ScienQuest platform. This platform was initially created for the study of the Scientext scientific texts corpus. This integration takes into account the metadata to the TALN corpus, and was done in an effort to reconcile these two corpora's sets of labels, and to convert for the TALN corpus the predefined queries designed for the Scientext corpus, in order to easily perform similar queries on the two corpora.

Mots-clés : corpus, corpus arborés, environnement d'étude de corpus.
Keywords: corpora, treebanks, corpus study environment.

1 Introduction

Les corpus de textes disciplinaires permettent d'étudier, en diachronie, l'historique d'une discipline et les évolutions de sa phraséologique, et d'un point de vue synchronique, de comparer les différents types de communications au sein de la discipline, mais aussi, pour peu que l'on dispose des corpus correspondants, par rapport à d'autres disciplines. Le corpus des actes de TALN offre ainsi l'opportunité à la communauté d'effectuer un peu d'introspection. Cette démonstration présente l'intégration de ce corpus dans la plateforme ScienQuest[1], qui intègre déjà un corpus pluridisciplinaire de textes scientifiques[2], et permet d'effectuer simplement des recherches et comparaisons sur des corpus arborés.

2 Le corpus TALN

Le corpus « 'TALN Archives »[3] a été collecté par Florian Boudin (Boudin, 2013) à partir des différents sites Web des conférences TALN et RÉCITAL (1997-2014). Il s'agit d'un corpus de textes au format *pdf*, accompagnés de méta-données (notice *bibtex* et résumé).

Un sous-ensemble de 586 articles a été sélectionné et traité par Ludovic Tanguy (Tanguy, 2013), afin d'en extraire le texte intégral, et de l'analyser avec TALISMANE (Urieli & Tanguy, 2013). Le corpus arboré ainsi obtenu contient 2,3 millions de tokens, annotés en parties du discours, en lemmes et en dépendances syntaxiques.

---

1 http://corpora.aiakide.net/link/taln
2 http://corpora.aiakide.net/link/sctexts-fr
3 https://github.com/boudinfl/taln-archives

3 Intégration à la plateforme ScienQuest

La plateforme ScienQuest (Falaise *et al.*, 2011) fut initialement créée pour l'étude linguistique du positionnement et du raisonnement dans le corpus de textes scientifiques Scientext (Tutin *et al.*, 2009), analysé avec Syntex. Cette plateforme, qui se veut simple à utiliser pour des non-TAListes, permet de rechercher en ligne des concordances dans un corpus, en fonction de critères linguistiques.

Méta-données. L'interface de ScienQuest permet de créer facilement des sous-corpus en fonction des méta-données du corpus. En outre, lors d'une recherche de concordances, des statistiques concernant la répartition des occurrences en fonction de ces méta-données sont calculées. Certaines des méta-données présentes dans le corpus TALN ont ainsi été intégrées dans la plateforme : conférence (TALN ou RECITAL), année et type d'article (court, long, etc.). Cela permet ainsi par exemple de sélectionner un sous-corpus en fonction de la conférence, ou de distinguer la fréquence relative d'un token ou d'un motif en fonction de l'année.

Étiquettes. Le jeu d'étiquettes utilisé par TALISMANE est assez riche (28 étiquettes morphosyntaxiques et 23 relations syntaxiques). Le mode de recherche privilégié dans ScienQuest (« mode libre ») utilise un assistant, qui présente ces étiquettes de manière conviviale. En général, le jeu d'étiquettes qui y est présenté est plus simple que le jeu d'étiquettes du corpus, et regroupe souvent plusieurs étiquettes sous un même nom de manière transparente pour l'utilisateur, qui ne voit que la « méta-étiquette » ; le jeu d'étiquettes complet n'est disponible dans sa totalité que dans le « mode avancé ». Pour le corpus TALN, afin de permettre une certaine compatibilité avec le corpus de textes scientifiques Scientext, nous nous sommes alignés, dans la mesure du possible, sur les 9 étiquettes morphosyntaxiques et les 13 relations syntaxiques présentées dans l'assistant « mode libre » pour ce corpus.

Grammaires. Enfin, le corpus Scientext est accompagné de « grammaires », c'est à dire de requêtes complexes pré-enregistrées, visant l'étude de la phraséologie du raisonnement et du positionnement dans les textes scientfiques. Ces grammaires, développées pour un corpus analysé avec Syntex, ont été « traduites » pour correspondre à l'analyse TALISMANE du corpus TALN.

4 Conclusion

L'intégration du corpus TALN dans ScienQuest permet ainsi d'effectuer facilement des recherches sur ce corpus. Cela peut aller de requêtes très simples, comme l'observation de l'évolution du nombre d'occurrences de « corpus » ou « grammaire » au fil des années, à des requêtes plus complexes, par exemple en utilisant les grammaires pré-enregistrées. On peut ainsi remarquer que la phraséologie semble assez homogène entre TALN et RÉCITAL, mais varie nettement en fonction du type d'article.

Références

BOUDIN F. (2013). TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue. *Actes de TALN* 2013, Les Sables d'Olonne, pages 507-514.

FALAISE A., TUTIN A., KRAIF O. (2011). Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques. *Actes de TALN* 2011, Montpellier, pages 187-215.

TANGUY L. (2013). Corpus TALN, en ligne : http://redac.univ-tlse2.fr/corpus/taln.html (consulté le 8 mai 2015).

TUTIN A., GOSSMANN F., FALAISE A., KRAIF O. (2009). Autour du projet Scientext : étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques. *Journées Linguistique de Corpus*, Lorient.

URIELI A., TANGUY L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. Actes de *TALN* 2013, Les Sables d'Olonne, pages 188-201.