Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques

Achille Falaise (1), Agnès Tutin (2), Olivier Kraif (2)

(1) GETALP-LIG
(2) LIDILEM
achille.falaise@imag.fr, agnes.tutin@u-grenoble3.fr, olivier.kraif@u-grenoble3.fr

Résumé

L'exploitation de corpus analysés syntaxiquement (ou corpus arborés) pour le public non spécialiste n'est pas un problème trivial. Si la communauté du TAL 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 » développé(e)s dans le cadre du projet Scientext, qui met à disposition un corpus d'écrits scientifiques interrogeable par partie textuelle, par partie du discours et par fonction syntaxique. Les modes simples sont décrits : un mode libre et guidé, où l'utilisateur sélectionne lui-même les éléments de la requête, et un mode sémantique, qui comporte des grammaires locales préétablies à l'aide des fonctions syntaxiques.

Abstract

The exploitation of syntactically analysed corpora (or treebanks) by non-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 able to handle advanced queries. In this paper, we present queries methods for the general public developed during the Scientext project, which provides a searchable corpus of scientific texts searchable from textual part, part of speech and syntactic relation. The simple query modes are described: a guided query mode, where the user easily selects the elements of the query, and a semantic mode which includes local pre-established grammars using syntactic functions.

Mots-clés : environnement d'étude de corpus, corpus étiquetés et arborés, création de grammaires assistée, visualisation d'information linguistique
Keywords: corpus study environment, treebanks, assisted grammars creation, visualization of linguistic information

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 TAL, 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 ANR Scientext (http://scientext.msh-alpes.fr), qui propose des modes de recherche simples pour non spécialistes 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 KWIC (*Key Word In Context*). 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 (recherche libre guidée et recherche sémantique), et effectuons enfin un premier bilan de son utilisation après quelques mois d'existence publique qui révèle une nette préférence des utilisateurs pour les modes de recherche simples.

2 Des annotations linguistiques riches mais des outils de recherche encore trop complexes pour le non spécialiste

2.1 Les corpus arborés du français

Peu de corpus arborés sont actuellement disponibles pour le français. Le Corpus arboré du français (*French Treebank*) (Abeillé *et al.* 2003)[1], est un corpus d'un million de mots de textes journalistiques, annotés en constituants, disponible pour des travaux de recherche mais non consultable en ligne. Parmi les corpus arborés du français consultables en ligne, il n'existe, à notre connaissance qu'un outil, Corpuseye, développé par (Bick, 2005) dans le cadre du projet VISL et qui propose une analyse syntaxique dans le cadre de la grammaire de contrainte (*Constraint grammar*)[2]. Ce projet donne accès à de nombreux corpus arborés dans 12 langues européennes, tous traités syntaxiquement. Ces corpus sont interrogeables en ligne grâce à un environnement d'étude, qui se base toutefois sur un langage de requêtes complexe, et apparaît donc difficilement utilisable par un non-spécialiste.

2.2 Les environnements de recherche sur corpus

Il existe plusieurs environnements d'étude de corpus annotés linguistiquement, basés sur des langages de requête. Parfois, un environnement graphique facilite la formulation des requêtes. Si ces formulations sous mode graphique clarifient la syntaxe du langage, elles en conservent néanmoins généralement toute la capacité expressive. Cependant, même en mode graphique, ces environnements restent difficiles à « apprivoiser » pour l'utilisateur non spécialiste, et peut demander un investissement qu'il ne jugera pas acceptable.

La majorité des outils d'étude de corpus ne traitent pas le niveau syntaxique. Ainsi le *Corpus Query Processor* (CQP) (Christ, 1994), développé à l'*Institut für Maschinelle Sprachverarbeitung* de Stuttgart, qui est devenu un standard dans la communauté du TAL, permet de faire des recherches en ligne[3] dans des corpus étiquetés et lemmatisés, mais ne prend pas en charge les corpus arborés. Une interface graphique a existé, mais n'est plus maintenue à ce jour. Il faut donc passer par un langage à base d'expressions régulières, qui reste complexe pour un non-spécialiste. En revanche, l'interface du corpus étiqueté et lemmatisé Elicop[4](Mertens, 2002) nous semble particulièrement intéressante et a été une source d'inspiration pour notre interface. Elle est basée sur un formulaire facile à remplir ne nécessitant pas de connaître un langage de requêtes (*cf.* figure 1). Le système ne permet toutefois pas de restreindre le corpus d'étude; il est en outre limité à une recherche sur quatre mots et, là encore, les relations syntaxiques ne sont pas prises en compte.

Tigersearch[5] (Lezius, König, 2000) est l'un des seuls environnements graphique (mais hors ligne) permettant d'interroger des corpus arborés (de type syntagmatique) mais l'outil n'est plus maintenu à l'heure actuelle.

---

1 http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php

2 http://corp.hum.sdu.dk/

3 http://www.ims.uni-stuttgart.de/projekte/CQPDemos/cqpdemo.html

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

5 http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/

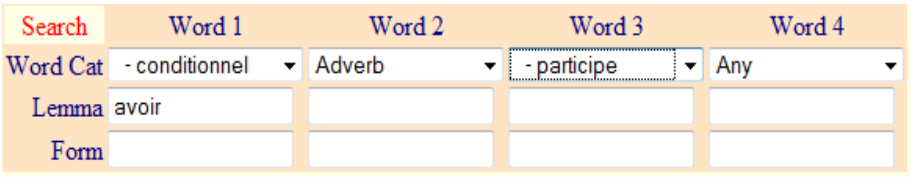

Figure 1: L'interface de requête du projet Elicop :
recherche sur le verbe avoir au conditionnel suivi d'un adverbe et d'un participe

En conclusion, nous ne pouvons que déplorer le manque d'environnements en ligne conviviaux pour la recherche dans des corpus annotés en français, en particulier pour les corpus arborés. Cette lacune est l'une des raisons qui a poussé à la réalisation de l'environnement Scientext, destiné à l'étude linguistique des écrits scientifiques.

3 Fonctionnalités de Scientext

L'utilisation de Scientext se compose de trois grandes étapes : choix d'un corpus, recherche de phénomènes linguistiques, et enfin affichage KWIC. Les exemples donnés ci-après sont basés sur le corpus de textes scientifiques français du projet Scientext, un corpus arboré traité avec l'analyseur Syntex de (Bourigault, 2007).

3.1 Sélection d'un sous-corpus

Par défaut, le système travaille sur un corpus entier. La première étape consiste soit à simplement accepter de travailler sur la totalité du corpus, soit à sélectionner un sous-corpus. Il est possible de combiner des groupes de textes préétablis suivant différents critères (par exemple discipline et type de texte scientifique). Les textes du corpus disposent d'annotations structurales (découpage en parties, titres, etc.), et il est également possible de restreindre la recherche à l'un de ces éléments. Une fois le corpus sélectionné (c*f.* exemple figure 2), l'utilisateur peut ensuite affiner la sélection en excluant certains textes.

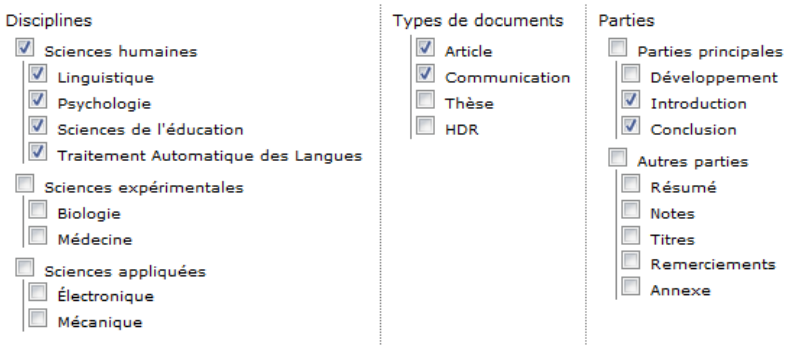

Figure 2: Un exemple de sélection du corpus Scientext : les introductions
et les conclusions des articles et communications en sciences humaines

3.2 Fonctionnalités de recherche

Une fois le corpus délimité, l'utilisateur est invité à choisir entre trois modes de recherche (c*f.* figure 3). Quel que soit le mode choisi, le résultat sera une grammaire locale, dont nous présenterons le langage plus loin dans ce papier. Cette grammaire locale est compilée vers le langage de requête utilisé par le moteur de recherche ConcQuest, développé par (Kraif, 2008), qui effectue la recherche dans le corpus.

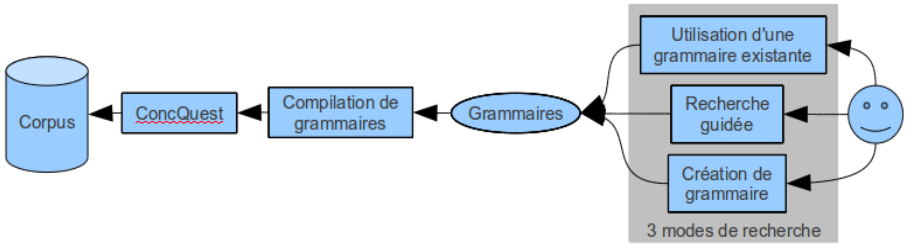

Figure 3: Architecture des modes de recherche dans Scientext

3.2.1 Recherche sémantique : utilisation d'une grammaire locale préétablie

Un ensemble de grammaires locales a été élaboré pour permettre une recherche sémantique dans les textes, de façon à ce que l'utilisateur n'ait pas à se soucier de la syntaxe complexe des requêtes. L'utilisateur choisit ainsi un premier type sémantique, par exemple l'évaluation et l'opinion, puis dans un second temps, comme dans la copie d'écran ci-contre (figure 4), un type spécifique les adjectifs d'opinion.

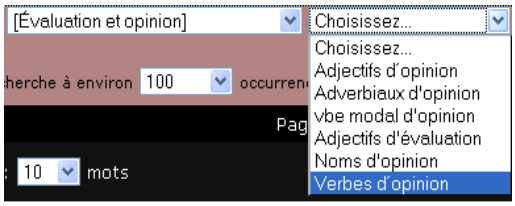

Figure 4: recherche sémantique

Une quinzaine de grammaires locales a été développée, principalement autour du thème du positionnement, thème privilégié dans le cadre du projet Scientext (Tutin *et al*. 2009). Nous envisageons de développer davantage de grammaires locales autour d'autres thèmes, en particulier dans une perspective didactique, pour la recherche d'expressions stéréotypées correspondant à des fonctions sémantiques et discursives, comme l'expression de la cause, le positionnement par rapport à d'autres auteurs, la formulation des problématiques.

3.2.2 Recherche guidée : un assistant pour effectuer des recherches libres

Figure 5 : état initial de la recherche guidée

En mode recherche guidée, l'interface se présente d'abord de manière minimaliste, avec un champ de saisie pour une seule contrainte sur un seul mot (*cf*. figure 5). Des boutons permettent d'ajouter des mots et des contraintes sur les formes, les lemmes, les parties du discours (et éventuellement des sous-catégories). Les expressions régulières sont acceptées. En l'absence de relations syntaxiques, l'ordre des mots dans le formulaire est pris en compte lors de la recherche.

Lorsqu'au moins deux mots sont présents, la possibilité est offerte de spécifier une relation syntaxique entre ces mots. Si une relation est choisie, l'ordre des mots n'est alors plus pris en compte (*cf*. figure 6).

Lors de la recherche, le contenu du formulaire est automatiquement converti en grammaire locale. Le mode guidé permet d'effectuer des recherches suffisamment complexes pour la plupart des utilisateurs. Il est volontairement limité, afin de ne présenter qu'un sous-ensemble facilement compréhensible de fonctionnalités utilisables de façon intuitive, sans recourir à une documentation. Pour exploiter toute l'expressivité de l'outil de recherche, il faut passer en mode de recherche avancée.

Figure 6: recherche guidée pour deux mots reliés par une relation syntaxique

3.2.3 Recherche avancée : un langage de requêtes pour les corpus arborés

Le mode recherche avancée permet de créer directement une grammaire locale, en suivant la documentation fournie. Ce mode est évidemment destiné aux utilisateurs spécialistes, c'est pourquoi nous n'en présentons dans cet article que les principales caractéristiques. Le langage de grammaire locale permet de spécifier des contraintes sur les mots (forme, lemme, partie du discours, flexion), un ordre entre mots, et des relations syntaxiques entre mots. Il est aussi possible de spécifier des listes de mots et des variables.

Certaines fonctionnalités innovantes sont spécifiques au traitement des corpus arborés, en particulier la possibilité d'étendre les relations syntaxiques présentes dans le corpus. Par exemple, l'analyseur Syntex effectue une analyse syntaxique de surface et ainsi ne crée pas de relation de dépendance directe entre un verbe à un temps composé et son sujet ; mais crée à la place une relation SUJ (sujet) entre le sujet et l'auxiliaire, et une relation AUX (auxiliaire) entre

l'auxiliaire et le verbe. Dans Scientext, il est possible de définir une relation « sujet » générique qui prend en compte ce cas de figure. La grammaire ci-dessous présente ainsi une grammaire détectant les verbes d'opinion et leur sujet en syntaxe profonde (par exemple la relation entre *je* et *penser* dans *nous avons pensé, nous pouvons penser*).

// Verbes d'opinion se rapportant à l'auteur, selon le schéma : verbe_opinion –(agent)-->auteur
// Ex : Je pense, j'ai considéré, nous pouvons penser

(SUJINF,#2,#1) = (SUJ,#3,#1)(OBJ,#3,#2) *// Pour traiter les structures avec infinitives*
(SUJCOMP,#2,#1) = (SUJ,#3,#1) (AUX,#3,#2) *// Pour traiter les structures avec auxiliaires*
(SUJGENERIQUE,#2,#1) = (SUJINF,#2,#1) OR (SUJ,#1,#2) OR (SUJCOMP,#2,#1) *// Relation "Sujet" générique*
$pron_auteur = nous, je, on *// Liste de pronoms pouvant désigner l'auteur*
$v_opinion = adhérer, admettre, adopter, affirmer, avancer, considérer, contredire, convenir, critiquer, croire, défendre, dénoncer, douter, espérer, estimer, juger, justifier, penser, postuler, préférer, privilégier, reconnaître, récuser, réfuter, regretter, rejeter, souhaiter, souligner, souscrire, soutenir, suggérer *// Liste de verbes d'opinion*
Main = <lemma=$v_opinion,#1> && <lemma=$pron_auteur,#2> :: (SUJGENERIQUE,#1,#2) *// Règle principale*

3.3 Visualisation des résultats

Les résultats d'une recherche sont ensuite consultables en affichage KWIC (figure 7), qui intègre les informations sur le type de texte. L'utilisateur peut en outre désactiver les résultats incorrects, qui ne seront pas extraits par la suite. Ces résultats sont exportables en CSV et en HTML. Il est possible d'élargir le contexte pour une ligne donnée ; le texte source est alors affiché, en respectant, lorsque l'information est disponible, le style du texte original (paragraphe, italique, *etc.*). Les dépendances syntaxiques de la phrase affichée peuvent être visualisées.

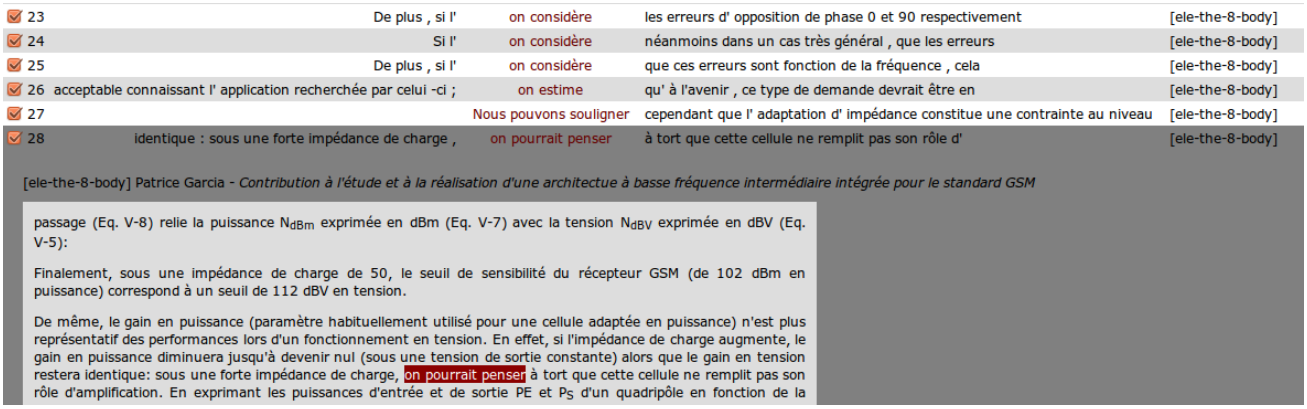

Figure 7: visualisation des résultats

Des statistiques sur les occurrences trouvées sont disponibles (*cf.* figure 8), par exemple le nombre d'occurrences et le pourcentage des lemmes et des formes et leur distribution par discipline, genre textuel, partie textuelle, et par texte. Il s'agit d'un type de fonctionnalité encore peu présent dans les outils d'étude de corpus, particulièrement intéressant pour l'étude des structures rhétoriques dans l'écrit scientifique.

Partie textuelle	Nombre absolu d'occurrences		Nombre de mots total		Nombre relatif d'occurrences
Développement	2112	/	4086783	=	5.17 ‰
Notes	131	/	255286	=	5.13 ‰
Introduction	104	/	223732	=	4.65 ‰
Conclusion	70	/	91272	=	7.67 ‰
Remerciements	37	/	21916	=	16.88 ‰
Annexe	34	/	116597	=	2.92 ‰
Résumé	7	/	27342	=	2.56 ‰

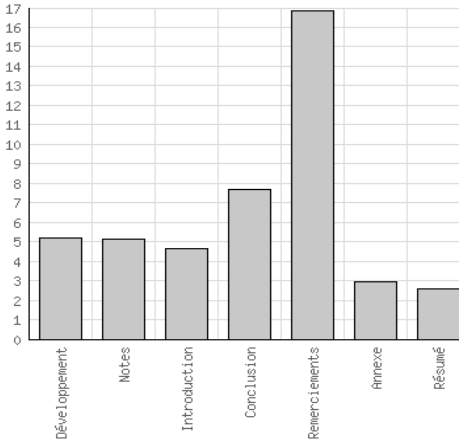

Figure 8: répartition des expressions de l'opinion de l'auteur dans le corpus suivant la partie textuelle (ci-contre : visualisation graphique)

4 Premier bilan et conclusion

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 (http://scientext.msh-

alpes.fr/fuls/), et intègre de nouveaux corpus, traités avec un analyseur différent, pour le projet ANR EMOLEX (http://scientext.msh-alpes.fr/emolex/).

Sur la période allant du début du lancement public du site fin juin 2010 à aujourd'hui (début avril 2011), 4432 requêtes ont été effectuées (en 662 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, en particulier sur les fonctions (exemple : sujet, objet, épithète) 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, pour le mode guidé l'ajout de nouvelles fonctionnalités (par exemple une préselection des relations syntaxiques à partir des parties du discours choisies). Ces améliorations seront testées sur un ensemble d'utilisateurs non spécialistes. L'utilisation des outils de TAL par les non spécialistes ne pourra pas faire l'économie d'une réflexion sur l'ergonomie de ces outils.

Références

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

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

Bick Eckhard (2005). Live use of Corpus data and Corpus annotation tools in CALL: Some new developments in VISL. Henrik Holmboe (ed.), *Nordic Language Technology, Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004* (Yearbook 2004), pp.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. Actes de *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, pp. 625-634. Lyon: Presses universitaires de Lyon.

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

Mertens Piet (2002). Les corpus de français parlé ELICOP : consultation et exploitation. Jean Binon, Piet Desmet, Jan Elen, Piet Mertens, Lies Sercu (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. S. Koeva, D. Maurel, M. Silberztein (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.