Analyse discursive pour le repérage automatique de segments obsolescents dans des documents encyclopédiques.

Marion Laignelet

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'université Toulouse II-Le Mirail

Discipline : Linguistique - Informatique

Présentée et soutenue par Marion LAIGNELET

Le 25 septembre 2009

Titre : Analyse discursive pour le repérage automatique de segments obsolescents dans des documents encyclopédiques.

JURY :

Liesbeth DEGAND	rapporteure	*Université de Louvain, Belgique*
Patrice ENJALBERT	rapporteur	*Université de Caen*
Agnès TUTIN	examinatrice	*Université de Grenoble 3*
Claude DE LOUPY	examinateur	*Laboratoire Syllabs, Paris*
Marie-Paule PÉRY-WOODLEY	directrice	*Université de Toulouse 2 - Le Mirail*
Ludovic TANGUY	encadrant	*Université de Toulouse 2 - Le Mirail*

École doctorale : CLESCO

Unité de recherche : Laboratoire CLLE-ERSS

Laboratoire Cognition Langues Langages Ergonomie

Équipe de Recherche en Syntaxe et Sémantique

L'ensemble de la thèse ainsi que les fichiers sources sont disponibles à l'adresse suivante : `http://marion.laignelet.free.fr`.

Concernant les ressources et outils informatiques créés, ils sont également disponibles à l'adresse sus-mentionnée, sous licence LGPL (`http://www.linux-france.org/article/these/licence/lgpl/lgpl_monoblock.html`).

Résumé

La question de la mise à jour des documents se pose dans de nombreux domaines. Elle est centrale dans le domaine de l'édition encyclopédique : les ouvrages publiés doivent être continuellement vérifiés afin de ne pas mettre en avant des informations fausses ou altérées par le temps.

Dans ce travail, nous proposons la mise en œuvre d'un prototype d'aide à la mise à jour : l'objectif visé est le repérage automatique de zones textuelles dans lesquelles l'information est potentiellement obsolescente.

Pour y répondre, nous proposons la prise en compte d'indices linguistiques et discursifs variés et faisant appel à des niveaux d'analyses différents. L'obsolescence étant un phénomène non linguistique pour lequel il n'existe pas d'outil rhétorique dédié, notre hypothèse est qu'il faut considérer les indices linguistiques et discursifs en termes de complémentarité, de combinaisons.

Sur un corpus annoté manuellement par des experts, nous projetons un repérage automatique d'un grand nombre d'indices linguistiques, discursifs et structurels. Un système d'apprentissage automatique est ensuite mis en place afin de faire émerger les configurations d'indices pertinentes dans les segments obsolescents caractérisés par les experts.

Notre double finalité est remplie : nous proposons une description fine de l'obsolescence dans notre corpus de textes encyclopédiques ainsi qu'un prototype logiciel d'aide à la mise à jour des textes.

Une double évaluation a été menée : par validation croisée sur le corpus d'apprentissage et par les experts sur un corpus de test. Les résultats sont encourageants. Ils nous amènent à faire évoluer la définition du segment d'obsolescence, sur la base des « découvertes » émergeant des corpus et dans l'interaction avec les besoins des experts concernant l'aide à la mise à jour. Ils montrent également les limites des annotations automatiques des indices linguistiques et discursifs.

Enfin, la reproductibilité de notre système doit être évaluée ainsi que la pertinence et la réutilisabilité du modèle de représentation des données présenté.

Abstract

The question of document updating arises in many areas. It is central to the field of encyclopedia publishing : encyclopedias must be constantly checked in order not to put forward wrong or time-altered information. In this study, we describe the implementation of a prototype of an aid to updating. Its aims is to automatically locate zones of text in which information might be obsolescent.

The method we propose takes into account various linguistic and discursive cues calling on different levels of analysis. As obsolescence is a non-linguistic phenomenon for which no specific rhetorical tool exists, our hypothesis is that linguistic and discursive cues must be considered in terms of complementarity and combinations.

Our corpus is first manually annotated by experts for zones of obsolescence. We then apply automatic tagging of a large number of linguistic, discursive and structural cues onto the annotated corpus. A machine learning system is then implemented to bring out relevant cue configurations in the obsolescent segments characterized by the experts.

Both our objectives have been achieved : we propose a detailed description of obsolescence in our corpus of encyclopaedic texts as well as a prototype aid to updating.

A double evaluation was carried out : by cross validation on the corpus used for machine learning and by experts on a test corpus. Results are encouraging. They lead us to an evolution of the definition of obsolescent segments, first, on the basis of the "discoveries" emerging from our corpora and also through interaction with the needs of the experts with respect to an aid to updating. The results also show limits in the automatic tagging of the linguistic and discursive cues.

Finally, the reproducibility of our system must be evaluated as well as the relevance and reusability of the model of data representation.

Remerciements

À mon jury : Liesbeth Degand, Patrice Enjalbert, Agnès Tutin, Claude de Loupy.

À Marie-Paule Péry-Woodley qui m'a fait confiance toutes ces années, et à Ludovic Tanguy qui a fait preuve de tant de patience.

À Didier Bourigault et Eric Marson, sans qui ce projet n'aurait probablement pas vu le jour.

À Frédéric Gardes qui au quotidien m'a encouragée, soutenue et toujours poussée vers l'avant, qui m'a patiemment aidé à comprendre les statistiques et la modélisation.

À François Rioult qui m'a donné les moyens et la technique en apprentissage automatique pour finaliser mon projet.

À Philippe Pleuvret pour sa confiance et ses conseils en statistiques.

À l'équipe du département des Encyclopédies Larousse : Yves Garnier, Jacques Florent, Marion Pépin, Mady Vinciguerra, Philippe Lacrouts, Line Karoubi et Pierre Chiesa.

À tous ceux qui ont eu le courage de relire les chapitres de cette thèse et de m'apporter de si bons conseils : Frédérik Bilhaut, Andrée Borillo, Cécile Fabre, Mai Ho-Dac, Marie-Paule Jacques, Franck Sajous.

À Christine Pernet, à l'écoute de mes répétitions de soutenance.

À l'ERSS et plus largement à CLLE-ERSS et CLLE-LTC : Anne Le Draoulec, Nabil Hathout, Myriam Bras, Fabio Montermini, Jean-François Bonnefon, Michel Aurnague, Anne Condamine, Jesse Tseng, Anna Kupsc, Pascale Vergely, Anne Przewozny, Josette Rebeyrolles, Jean-Michel Tarrier, Philippe Muller, Laure Sarda, Marianne Vergez, Clémentine Adam, Julien Eychenne, Aurélie Picton, Anne-Lise Coquillon, Béatrice-Akissi Boutin, Nathalie Dehaut, Buddy Dirat, Aurélie Guerrero, Edith Galy, Christophe Pimm, Stéphanie Lopez, Sylvain Navarro, Kuna Mvogo, Annique Smeding, Maria Fohlin.

À toutes les rencontres faites ici et là au détour d'une conférence, d'un colloque, d'un projet : Antoine Widlöcher, Marie Chagnoux, Stéphane Ferrari, Yann Mathet, Guy Lapalme, Nathalie Aussenac-Gilles, Mauro Gaio, Aïssa Derrouaz, Amanda Bouffier, Thierry Fontenelle.

À Corine Prunier, Laurence Lamy, Nathalie Moulic, Corine Ratier, Bruno Chenu et Geneviève Usache.

À ma famille, Fred, Lena et Asalais qui m'ont permis de penser à autre chose

qu'à la linguistique et à l'informatique (jusqu'à 18 couches par jour pendant 6 mois, ca aide pour penser à autre chose...).

À mon père pour ses remarques et re-lectures, son implication dans mes travaux.

À ma mère pour m'avoir appris à toujours aller de l'avant, pour sa tendresse et son dévouement.

À mon pépé pour tout son amour. Et une douce pensée pour ma Granny qui continue chaque jour à veiller sur moi.

À Anne et Charles, ma chère fratrie, toujours encourageante et si fière de moi.

À Tatalaso, ma belle-soeur survoltée, à Cathy, belle maman à l'écoute et à Michel, beau-papa dynamique.

À tous mes amis : Anne-Marie et Julien, Benoît et Audrey, Gilles, Aurélie, Virginie, Caroline, Sandra et Benoît, et toutes les mamans (et futures mamans) qui m'ont aidée et soutenue.

Table des matières

Liste des tableaux

Table des figures

Introduction

De l'homme à la machine : la mise à jour des encyclopédies

Le monde de l'édition et plus spécifiquement celui de l'édition encyclopédique connaît actuellement une réelle mutation technique. Le nombre de mises en ligne d'encyclopédies généralistes sur Internet en est révélateur. Ainsi, *Wikipédia* est lancé en 2001[1] sur Internet, *Larousse* en 2008[2] ; *Hachette*[3], *Encyclopedia Universalis*[4], *Encarta*[5], ou encore *Encyclopédie gratuite*[6] sont également disponibles sur Internet.

Face à cette réalité, les acteurs de l'édition doivent être les plus réactifs possible aux évolutions des connaissances dans tous les domaines qu'ils couvrent. Il est toutefois difficile d'envisager une mise à jour systématique et perpétuelle de l'information dans les encyclopédies du fait de la taille de ce type de documents. Et à l'heure où Internet permet un accès permanent à l'information, il est important pour les éditeurs d'être également capables de proposer une mise à jour en temps réel des informations diffusées au public, tant pour leurs versions en ligne que pour les versions papier même si pour ces dernières le problème de la mise à jour est quelque peu différent, d'un point de vue temporel et éditorial du moins.

Les enjeux sont bien réels et il est facile de deviner les difficultés auxquelles sont confrontées les maisons d'édition lorsqu'elles souhaitent proposer de nouvelles versions de leurs encyclopédies : comment mettre à jour des contenus éditoriaux si importants en taille qu'ils peuvent atteindre vingt volumes de plusieurs centaines de pages chacun ? Et surtout comment le faire de la manière la plus exhaustive et le plus rapidement possible ?

Wikipedia a choisi la solution coopérative : il s'agit de laisser les internautes participer eux-mêmes à la rédaction des articles en ligne. Sur Internet, le choix des Éditions Larousse[7] est différent : la modification des articles originaux de l'ency-

[1] `http://fr.wikipedia.org/wiki/Accueil`

[2] `http://www.larousse.fr/encyclopedie/`

[3] `http://www.ehmelhm.hachette-multimedia.fr`

[4] `http://www.universalis.fr/`

[5] `http://fr.encarta.msn.com/`

[6] `http://www.encyclopedie-gratuite.fr/`

[7] Nous remercions chaleureusement les Éditions Larousse et plus précisément Yves Garnier, Jacques Florent, Line Karoubi, Philippe Lacrouts et Marion Pépin pour leur aide, leur confiance

clopédie n'est pas autorisée et il existe en parallèle un forum coopératif accessible et modifiable par tous. Que ce soit pour les versions papier ou les versions CD-Rom des Éditions Larousse ou le format de fascicule des Éditions Atlas[8], la question de la mise à jour est cruciale.

La possibilité de l'automatisation, de la création de dispositifs d'aide à la mise à jour est centrale pour un secteur en profonde mutation qui doit d'un côté être capable de satisfaire la demande (une information *à jour*) et de l'autre être à même de concurrencer efficacement des encyclopédies coopératives.

Le projet de recherche présenté dans cette thèse a pris forme au sein de l'entreprise d'édition-packaging[9], INITIALES dans le cadre d'un contrat CIFRE. La question de la mise à jour de données textuelles massives s'est posée lorsque cette société s'est trouvée en charge d'éditer, de *packager*[10] et de mettre à jour des fiches encyclopédiques publiées par les Éditions ATLAS. Il s'agit de fascicules, de fiches encyclopédiques thématiques envoyées aux clients sous forme d'abonnement par voie postale. Ce système est nommé *éditions au long cours*[11], elles durent généralement trois ans, voire plus (il ne s'agit pas d'éditer une fois un ouvrage mais d'étaler les éditions et publications des fiches sur un temps déterminé).

Aux éditions LAROUSSE, la question de la mise à jour se pose depuis longtemps. Plusieurs stratégies ont été envisagées, la stratégie actuelle consistant à demander aux auteurs de mettre le moins possible d'informations susceptibles d'évolutions. Une stratégie intéressante a été mise en place : il était demandé aux auteurs de se prononcer directement au moment de la rédaction des articles sur la nature potentiellement évolutive des informations (par balisage XML). Mais cette solution a été abandonnée assez vite car elle s'est avérée contraignante pour les auteurs (temps supplémentaire). Aujourd'hui, les entrées des dictionnaires sont mises à jour de façon arbitraire et aléatoire, en fonction de l'actualité et des événements importants.

Que ce soit au sein des Éditions Atlas ou au sein des Éditions Larousse, les mises à jour des documents encyclopédiques se font sensiblement de la même manière.

Chez Initiales, trois niveaux de mise à jour sont recensés :
- niveau 1 : correction des captures d'écran/images et des légendes ;
- niveau 2 : correction des textes et des captures d'écran et des légendes ;
- niveau 3 : fiche quasiment réécrite en entier et correction des captures d'écran et des légendes.

Aux Éditions Larousse, il existe également trois niveaux de mise à jour :

(prêt des encyclopédies) et l'intérêt qu'ils portent à ce travail de recherche.

[8] Nous remercions chaleureusement Aïssa Derrouaz qui a eu l'extrême amabilité de faire le lien avec les Éditions Atlas. Nous n'oublions bien entendu pas Éric Marson, de la Société Initiales, qui est à l'origine de ce projet de thèse et qui nous a permis de travailler sur les données textuelles des Éditions Atlas.

[9] Le packaging concerne la mise en page, la mise en forme des livres.

[10] Un packager est chargé de la mise en forme des livres (intégration des textes, d'images, de graphisme, etc.).

[11] ou *collection vendue à tempérament*.

– niveau *nouvelle présentation* : au minimum, changement de la couverture, au maximum modification des couleurs ;
– niveau *édition mise à jour* : corrections ponctuelles des entrées du dictionnaire ;
– niveau *nouvelle édition* : refonte complète, grosse mise à jour des entrées.

La mise à jour des textes est entièrement manuelle. Chez Initiales, environ 60 personnes externes à la société travaillaient[12] à la rédaction et à la mise à jour des fiches[13]. Pour chacune des fiches est indiqué le niveau de révision (« 3^e révision », « 4^e révision », etc). Chez Larousse, la mise à jour se fait en interne pour les deux premiers niveaux, et par les auteurs des entrées pour le troisième niveau. D'une manière générale, les corrections se font selon l'actualité politique, sociale, scientifique, etc.

L'objectif industriel et la visée scientifique appliquée de cette thèse consistent en la création d'un prototype logiciel d'aide à la mise à jour de contenus encyclopédiques pour l'édition. Idéalement, l'appui logiciel concerne les mises à jour de niveaux 2 et 3 ou *édition mise à jour* et *nouvelle édition*. Concrètement, il s'agit de proposer à un utilisateur-rédacteur un outil de navigation intra-documentaire signalant le caractère potentiellement obsolescent des informations situées dans telle ou telle partie d'un document.

Intrinsèquement liée à la notion de mise à jour, celle de l'obsolescence va nous guider tout au long de ce travail. Nous allons maintenant décrire ce que ce terme recouvre.

L'obsolescence : quelle réalité pour l'édition ?

Le mot « obsolescence » vient du latin « tomber en désuétude ». La définition de ce terme par le Grand Robert s'inscrit dans le domaine technique :

> *« Vieillissement de l'équipement industriel, dû à l'apparition d'un matériel nouveau » (Le Grand Robert)*

Les aspects économiques et sociétaux sont également intrinsèquement liés à l'obsolescence ainsi que le montre la citation suivante :

> *« L'obsolescence a été étudiée et changée en technique. Les spécialistes de l'obsolescence connaissent l'espérance de vie des choses : trois ans, une salle de bain ; cinq ans, un living-room ; huit ans, un élément de chambre à coucher ; trois ans, l'aménagement d'un point de vente local, une auto, etc. »* (Henri Lefebvre, La vie quotidienne dans le monde moderne, in Le Grand Robert, p. 157)

Dans un sens, les encyclopédies actuelles répondent à cette définition : le client d'une encyclopédie (ou d'un dictionnaire) peut trouver sa collection obsolète car il

[12] La société n'existe plus à l'heure de la rédaction de cette thèse.

[13] Bien que la collection des *Mémofiche* soit terminée, le travail de correction des fiches se poursuit au sein des Éditions Atlas.

n'y trouve pas les derniers mots ou concepts à la mode ou nouvellement créés et utilisés. Il cherchera alors naturellement à s'en procurer un exemplaire plus récent, plus en adéquation avec ses attentes : chiffres et dates récentes, nouveaux mots, derniers événements politiques, économiques, sociaux, etc.

Mais une encyclopédie est un objet qui n'a pas la même fonction dans la société qu'un objet de la vie courante comme une salle de bain ou un living-room. Elle a pour but de permettre l'accès à une forme de savoir absolu et vrai pour tout citoyen. Pour rappel, les premières lignes écrite par Diderot en 1751 dans l'Encyclopédie :

> *« Le but d'une encyclopédie est de rassembler les connaissances éparses sur la surface de la terre ; d'en exposer le système général aux hommes avec qui nous vivons, et de le transmettre aux hommes qui viendront après nous ; afin que les travaux des siècles passés n'aient pas été inutiles pour les siècles qui succéderont ; que nos neveux devenant plus instruits, deviennent en même temps plus vertueux et plus heureux ; et que nous ne mourions pas sans avoir bien mérité du genre humain. »*

Dans une encyclopédie, l'obsolescence est un phénomène particulier qui ne remet pas en cause le bien-fondé de l'existence même de l'encyclopédie : il est en revanche indispensable de savoir le repérer et le corriger dans ce type d'ouvrage afin de parvenir à un juste équilibre entre des informations anciennes mais nécessaires à la compréhension du monde d'aujourd'hui et des informations qui ont pu évoluer et se modifier et qui sont susceptibles de rendre obsolète un objet (au sens économique du terme).

Une connaissance est difficilement évaluable en termes de vérité absolue ou non (vrai ou faux) même si cette connaissance porte sur un fait vérifiable du monde. Il est plus prudent de penser les connaissances d'une manière graduelle en considérant qu'elle sont, à un bout de l'échelle, certaines, et à l'autre bout de l'échelle, peu certaines. Entre ces deux extrêmes, une connaissance factuelle peut être soumise au doute, être jugée acceptable en attendant la preuve du contraire.

Dans ce contexte, nous considérons qu'une information obsolescente est une information qui n'est potentiellement plus vraie à $T + n$, qui peut être jugée douteuse ou subjective ou encore qui n'est plus *valable* à $T + n$ du fait de l'évolution temporelle ou technique. n correspond à une période temporelle définie arbitrairement : pour ce travail, nous considérons $T + 1\ an$ au minimum.

Le segment d'obsolescence : quelle réalité textuelle ?

De la présentation de notre projet découle l'idée qu'un *segment d'obsolescence* se définit d'abord par rapport à un usage concret, un besoin réel, à savoir la mise à jour éditoriale. Un segment d'obsolescence présente la particularité majeure de contenir une information dont la caractéristique est d'être susceptible d'évolution dans le temps. Ce segment textuel peut également être pertinent parce qu'il

véhicule des connaissances qui, relativement à des besoins éditoriaux, nécessiteraient d'être réactualisées.

L'exemple 1 présente un segment d'obsolescence. L'auteur y exprime une issue possible et probable concernant les recherches sur le sida.

1. Actualité

§ Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux. Toutefois, il convient de rappeler un certain nombre de découvertes très récentes. En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

1.1. Un vaccin contre le sida ?

§ Des recherches portant sur les prostituées [...]. La recherche se tourne justement aujourd'hui vers des vaccins qui [...]. Des expériences ont été faites pour [...]. En juin 2003, une équipe de biologistes américains a obtenu des résultats qui pourraient laisser envisager [...]. Les chercheurs sont parvenus [...]. Cette découverte pourrait aboutir à la mise au point d'un antigène [...]. [. . .]

Source : Corpus ATLAS (fiche Médecine - Le Sida)

Exemple 1 - *Un segment d'obsolescence*

La fiche d'où est extrait ce passage a été éditée et distribuée dans le courant de l'année 2003 par les Éditions Atlas. Ce type de fiche est destiné à être distribué dans le cadre d'éditions au long cours : ce sont des éditions qui fonctionnent sous forme d'abonnement ; un client s'abonne à un moment t (deux possibilités annuelles) et pendant une durée déterminée (par le client), il recevra un nombre déterminé de fiches tous les mois. Le début des abonnements est alors variable d'un client à l'autre, l'écart peut aller jusqu'à cinq ans : à cinq ans d'intervalle, une même fiche est potentiellement obsolète.

Un client qui s'abonne à l'encyclopédie en 2007 est susceptible de recevoir cette fiche, écrite en 2003 et dans laquelle cet exemple apparaît. Cependant, pendant ce laps de temps (de 2003 à 2007), soit certaines des prédictions formulées par l'auteur se seront réalisées, soit elles auront été repoussées par les scientifiques ou encore de nouvelles données peuvent entrer en jeu. Il est donc tout à fait souhaitable que ce segment de texte ait été préalablement mis à jour.

À travers cet exemple, nous pouvons observer l'importance de la composante temporelle. L'objectif de cette étude n'est ni de procéder à des calculs de la référence temporelle, ni de chercher à associer un événement particulier à une date particulière comme c'est le cas en extraction d'information ou dans les systèmes de question-réponse. Nous ne cherchons pas non plus à valider le fait que l'événement « une équipe de biologistes américains a obtenu des résultats qui [...] » est vrai, ni qu'il s'est réellement produit en « juin 2003 ».

Nous recherchons les segments pour lesquels il est pertinent de penser que l'information donnée est susceptible d'avoir évolué entre le moment de l'édition de la fiche et le moment de sa lecture potentielle par un client. Dans l'exemple 1, la phrase « Cette découverte pourrait aboutir à la mise au point d'un antigène » doit

être vérifiée et mise à jour. Le type d'information recherché peut être local (une date, un chiffre, un nom de société, etc.) ou au contraire à granularité variable, de la taille de la phrase à celle du paragraphe tout entier.

L'objectif de ce travail consiste à chercher et proposer des méthodes et techniques pour produire un *repérage* des segments textuels contenant une information potentiellement obsolescente. Il ne s'agit pas de se substituer aux rédacteurs des encyclopédies en effectuant à leur place la mise à jour des information. Nous proposons le développement d'outils s'intégrant dans un système d'*aide* à la mise à jour des textes encyclopédiques.

Travaux liés à la question de la mise à jour de l'information : rechercher l'expression du changement ou de l'évolution

Lors de nos recherches sur les études, applications ou logiciels en relation avec la tâche de mise à jour de contenus textuels, nous avons constaté qu'il y a assez peu de travaux sur cette question. C'est du domaine de la veille stratégique [14] que nos travaux semblent le plus proches.

La veille technologique est définie comme l'activité mettant en œuvre des techniques d'analyse d'informations sur un produit ou un procédé et sur l'état de l'art et l'évolution de son environnement scientifique, technique, industriel ou commercial. L'objectif est de collecter, organiser, puis analyser et diffuser les informations pertinentes qui vont permettre d'anticiper les évolutions, et qui vont aider à l'innovation. Techniquement, la veille stratégique fait appel principalement aux techniques de recherche d'information.

Un des objectifs de la veille stratégique est de rechercher les indicateurs de nouveauté et d'innovation : ce type d'informations permet à une entreprise de rester au fait des innovations technologiques. Les techniques mises en œuvre en veille stratégique sont relativement proches de celles mises en œuvre dans ce travail. Mais contrairement à la veille stratégique qui recherche les indices de nouveauté ou de changement, nous nous préoccupons de ce qui est de l'ordre de l'obsolescence.

Si nous n'avons pas trouvé de travaux sur la recherche de l'obsolescence au sens où nous l'entendons, il en existe sur les notions de nouveauté et de variation terminologique notamment.

Le travail de Ibekwe-SanJuan (2005) a pour objectif la recherche d'indices de nouveauté pour détecter ce qui change dans le contenu des textes eux-mêmes. L'auteur met en œuvre une méthodologie dans laquelle nous nous retrouvons notamment parce qu'elle fait appel à des indices linguistiques. L'auteur dit s'inspirer des travaux de Teufel *et al.* (1999). Elle travaille sur des textes courts (titres et résumés scientifiques) et oriente ses recherches sur l'expression de l'apport de l'auteur. Elle met en relief trois types d'indices : ceux qui rendent compte des *objectifs*,

[14] Ou scientifique ou technologique.

ceux qui montrent les *contributions/résultats* et enfin ceux qui apportent des *conclusions*. Ses travaux sont menés sur des textes en langue anglaise et les indices textuels qu'elle a mis en évidence sont des expressions du type « Here, we propose a novel (...) approach », « we discuss recent developments » (information apportée : la nouveauté), « Our research suggests that », « Results confirm that » (information apportée : résultats/contribution/conclusion), « In this article/paper/study, we examine/investigate/describe » (information apportée : objectif). À l'issue d'une observation manuelle des indices, elle crée des automates qui dans un nouveau corpus vont repérer automatiquement les types d'indices sus-cités. Les résultats semblent encourageants. L'auteur suppose que les indices qu'elle met en lumière sont potentiellement généralisables à des domaines scientifiques différents moyennant de légères variations. Elle reste malgré tout prudente quant au fait que toute nouveauté n'est pas systématiquement encadrée, marquée par des indices textuels et qu'il serait probablement pertinent de les coupler à des indices fréquentiels et temporels.

Les travaux de Condamines *et al.* (2004) concernent la question de l'évolution des connaissances à travers l'étude des termes au sein de domaines spécifiques (l'aérospatial). L'objectif est *« d'identifier et décrire les formes privilégiées du changement sémantique à partir de l'analyse linguistique d'un corpus spécialisé construit de façon à rendre possible l'observation d'évolutions de connaissances »*. Ces auteurs utilisent une méthode *outillée* par des traitements automatiques (extraction de termes, étiquetage grammatical, concordancier). Leur façon d'utiliser les corpus diffère cependant de la nôtre puisque leur choix a consisté à comparer les « mêmes » textes pris à des moments différents de leur évolution ce qui permet une réelle étude diachronique alors que notre étude est exclusivement synchronique (collecte de textes à un moment donné). Les travaux de Picton (2008) vont également dans ce sens : l'auteur recherche les indices linguistiques permettant d'accéder à l'évolution des connaissances sur des périodes courtes. Sur la base de l'étude de quatre types d'indices (fréquences, contextes d'évolution, variantes et dépendances syntaxiques), elle cherche à associer un ou plusieurs indices avec un ou plusieurs types d'évolution spécifique.

Hypothèses et méthodologie

L'obsolescence n'est pas une catégorie linguistique « normée », « traditionnelle » comme peuvent l'être les structures argumentatives par exemple. Il n'existe pas de classes lexicales ou syntaxiques exclusivement dédiées à l'obsolescence. Nous supposons que c'est sur la base d'une accumulation d'indices différents que l'interprétation d'un segment obsolescent est rendue possible.

Notre hypothèse est que les segments d'obsolescence peuvent être délimités sur la base de combinaisons d'indices linguistiques et discursifs. Ces indices ont une fonction *première* dans le discours, un rôle rhétorique, syntaxique ou lexical propre qui seul ne suffit pas pour repérer l'obsolescence. C'est en *combinaisons*

que les indices vont contribuer au repérage des segments obsolescents.

Le problème qui se pose alors est de *(i)* déterminer quels indices peuvent être de bons candidats en tant qu'éléments premiers des combinaisons, et *(ii)* déterminer ces combinaisons.

Pour déterminer les indices qui doivent être considérés, nous nous basons principalement sur l'étude fine du phénomène de l'obsolescence à partir d'un corpus annoté manuellement de l'obsolescence ainsi que sur nos intuitions linguistiques et discursives. Nous visons un nombre d'indice le plus large possible, prenant en compte des phénomènes linguistiques variés : à l'issue de cette étude exploratoire, il sera alors possible de préciser les indices ou classes d'indices pertinents pour l'obsolescence.

La méthodologie mise en œuvre dans ce travail s'apparente à l'approche *top-down* proposée par Biber *et al.* (2007)[15]. Il distingue sept étapes :

1. détermination des types d'unités discursives (catégorie fonctionnelle/communicative) : *pour nous, l'obsolescence* ;
2. segmentation des textes selon ces unités discursives (segmentation) : *pour nous, annotation manuelle par des experts-rédacteurs* ;
3. identification des types d'unités discursives (classification) : *exploration en corpus du phénomène* ;
4. analyse des caractéristiques des unités discursives : *analyse linguistique et discursive fine, repérage automatique des indices dans le corpus* ;
5. description des caractéristiques linguistiques typiques des unités discursives (description linguistique) : *analyse statistique sur le corpus, émergence des indices et combinaisons d'indices typiques* ;
6. description des structures discursives et textuelles (structure textuelle) : *émergence des relations/combinaisons d'indices typiques dans les segments d'obsolescence* ;
7. description des patrons génériques de l'organisation discursive (tendances organisationnelles discursives) : *évaluation de la pertinence des combinaisons d'indices*.

Biber *et al.* (2007) insistent sur le fait que cette méthodologie est fondée sur une analyse de type *corpus-based* : sur la base d'une collection de textes représentatifs pour un genre donné, les traits linguistiques et leurs interactions *émergent* de l'étude du corpus dans le cadre d'une hypothèse pré-établie (dans notre cas, notre hypothèse est que des configurations d'indices discursifs sont à même de délimiter les segments d'obsolescence). C'est également la voie méthodologique que nous suivons.

[15] Son objectif final est l'étude de la structure linguistique des textes sur la base d'une étude de traits lexico-syntaxiques et de leurs relations.

Plan de la thèse

L'élaboration du sujet de la présente thèse à partir de la demande appliquée s'insère dans un contexte de recherche alliant linguistique de corpus, Traitement Automatique des Langues (T.A.L.) et linguistique de discours.

Nous souhaitons montrer dans quelle mesure la prise en considération de connaissances linguistiques, et plus précisement d'une analyse discursive, est un gain pour la description et le repérage de ces segments particuliers contenant de l'information susceptible d'évoluer dans le temps. L'obsolescence est un phénomène complexe qui nécessite la prise en compte d'éléments linguistiques variés : le traitement du temps est probablement le plus évident, les expressions relevant de la subjectivité du rédacteur sont également importants. Nous exploitons des informations de type discursif (notamment à travers l'étude des titres) et liées à la structure du document (position des éléments dans le texte).

La prise en compte d'éléments linguistiques n'est pas nouvelle en T.A.L. et cette démarche s'inscrit dans une conception plus globale de certains systèmes de traitement de l'information. Dans le chapitre 1 (p. 15), une présentation des domaines et applications en traitement de l'information dans une visée T.A.L. est proposée. Ce chapitre est également l'occasion de situer ce travail parmi l'ensemble des applications existantes et des techniques utilisées actuellement que ce soit dans le monde professionnel ou dans le cadre de projets de recherche. Nous présentons également des travaux de recherche récents auxquels nous sommes liée, que ce soit théoriquement et/ou méthodologiquement.

L'obsolescence est un phénomène qui n'a pas encore été étudié en tant que tel et ni fait l'objet d'une recherche spécifique. Dans le chapitre 2 (p. 35), nous présentons la démarche d'exploration en corpus dont l'objectif est la description de ce phénomène. Nous avons mené cette phase exploratoire sur des textes fournis par les Édition Atlas et les Éditions Larousse. Ces corpus ont été annotés manuellement par des experts. À l'issue de ce processus, nous disposons d'un ensemble de phrases annotées selon leur caractère obsolescent ou non : ces textes sont à la base d'une caractérisation (quantitative et qualitative) fine de l'obsolescence et des segments d'obsolescence.

La seconde partie de ce mémoire est consacré à la caractérisation linguistique des segments d'obsolescence. Elle est basée sur l'exploration du corpus annoté manuellement : nos propos sont systématiquement illustrés d'exemples attestés issus de notre corpus de travail.

Le chapitre 3 (p. 65) propose une description détaillée des éléments linguistiques potentiellement marqueurs de l'obsolescence. Il s'agit notamment de montrer en quoi la composante énonciative du discours est importante pour ce travail. Nous présentons d'abord les aspects temporels, aspects intrinsèquement mêlés à l'obsolescence. Nous décrivons également les éléments exprimant le point de vue du rédacteur. D'autres éléments comme les expressions chiffrées, les lieux, les noms de personne sont également traités dans ce chapitre.

Le chapitre 4 (p. 89), présente plusieurs modèles et hypothèses de linguistique

du discours qui permettent d'envisager les textes selon des granularités variables. Après avoir fait un rappel sur les notions de texte, de discours, d'indice, de marqueur et de segment, nous présentons l'hypothèse de l'encadrement du discours, le modèle de l'architecture textuelle et enfin, la notion de prédiction.

La troisième partie du mémoire présente le dispositif expérimental mis en œuvre pour déterminer les indices potentiellement aptes à marquer l'obsolescence. Le schéma 2 (p. 11) résume les différentes étapes du dispositif expérimental (et également les chapitres de cette partie) : sur la base du corpus d'apprentissage annoté manuellement, les segments d'obsolescence sont caractérisés linguistiquement. Cette caractérisation linguistique (sémantique et discursive) constitue le fondement de l'outil ALIDIS[16] : il s'agit d'annoter automatiquement les textes d'indices linguistiques susceptibles d'être de bon marqueurs de l'obsolescence (chapitre 5). De ces annotations manuelles et automatiques est créée, à l'aide de l'outil OCAS[17], une représentation sémantique des textes permettant de s'abstraire de la linéarité des textes et de traiter statistiquement des indices linguistiques divers et variés (chapitre 6). Sur la base de cette abstraction sémantique, l'outil STAAT[18] (chapitre 7) met en œuvre des techniques statistiques (statistiques de base, Analyse en Composantes Principales et Apprentissage Automatique) dans le double objectif suivant : donner une meilleure description du phénomène de l'obsolescence et proposer des règles pour le repérage automatique des segments d'obsolescence dans les documents encyclopédiques.

Dans le chapitre 8, les observations issues de l'Analyse en Composantes Principales et de l'Apprentissage Automatique sont confrontées et mises en perspective par rapport aux travaux de T.A.L. exploitant des indices et des combinaisons d'indices linguistiques.

Le chapitre 9 présente le prototype d'aide à la mise à jour et son évaluation par des experts du domaine. Les règles apprises par le classifieur sont utilisées sur un corpus de test non annoté manuellement mais annoté pas l'outil ALIDIS (indices linguistiques et discursifs). Les règles d'apprentissage sont projetées et fournissent ainsi un prototype de système d'aide à la mise à jour des textes. Cette annotation automatique de l'obsolescence est finalement évaluée par des rédacteurs.

La méthodologie mise en place permet donc à la fois de proposer un prototype mais également de fournir une meilleure description des segments d'obsolescence et du phénomène en question. L'outil STAAT permet ainsi de réévaluer nos hypothèses de départ quant à la description linguistique de l'obsolescence.

[16] Pour : Analyse LInguistique des DIScours.

[17] Pour : Outil de Création d'Abstraction Sémantique.

[18] Pour : STatistiques et Apprentissage Automatique sur les Textes.

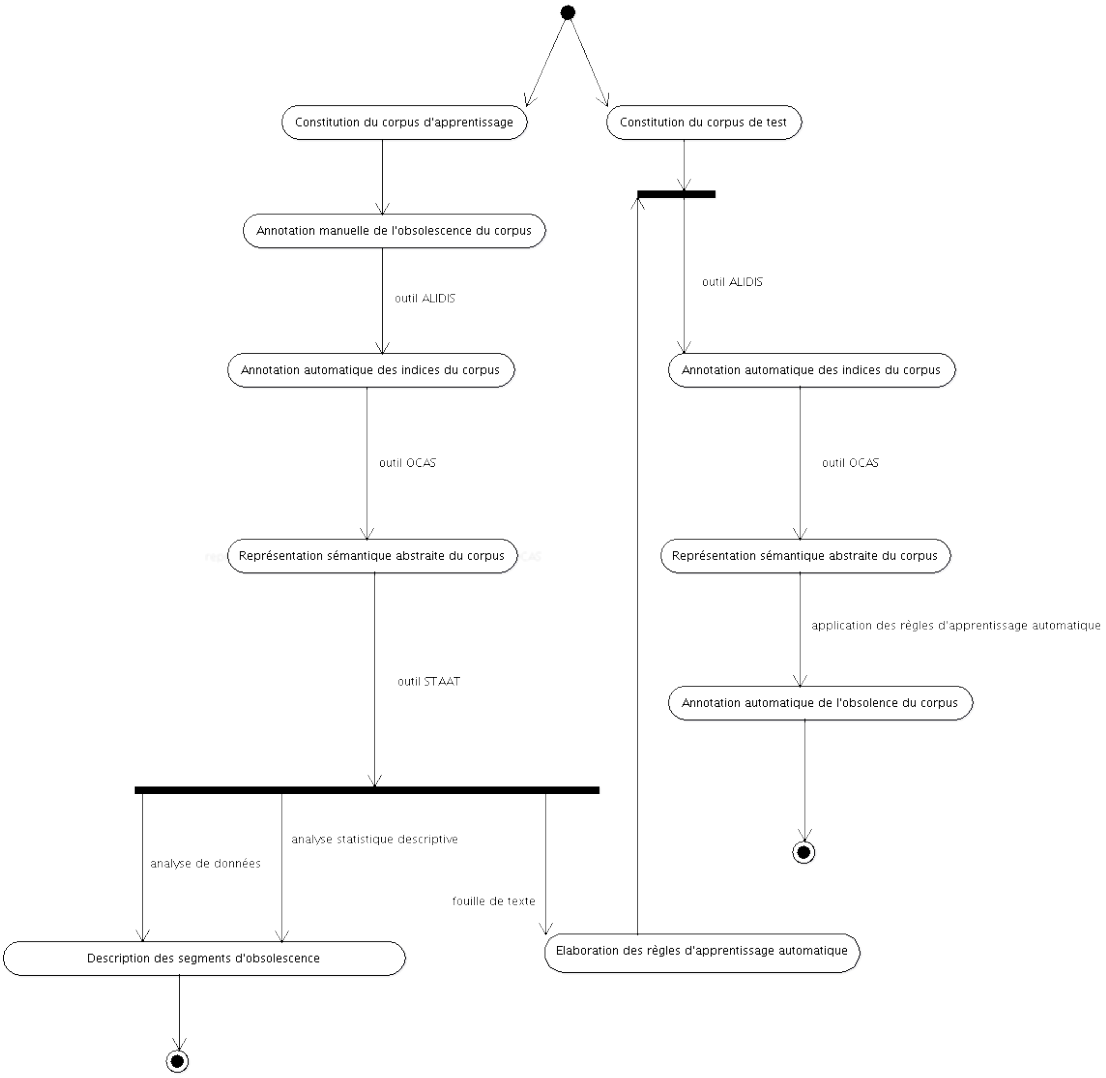


Schéma 2 - *Méthodologie générale pour le repérage automatique des segments d'obsolescence*

Première partie

Positionnement en T.A.L.

Chapitre 1

Applications et techniques actuelles en traitement de l'information

Ce chapitre propose un panorama des diverses applications en Traitement Automatique des Langues (T.A.L) au sein duquel ce travail s'inscrit. Seront d'abord introduites les applications de Recherche d'Information (grain d'analyse large) et d'Extraction d'Information (grain d'analyse fin) puis les systèmes faisant appel à la notion de grain d'analyse variable : cette conception granulaire est à la base de notre travail.

Un aperçu des techniques traditionnellement exploitées dans les systèmes de traitement de l'information est ensuite présenté. L'objectif est de montrer que le T.A.L. et la linguistique sont capables de fournir des méthodes adaptées à des besoins complexes lorsqu'il s'agit de traiter du « sens » des textes. À travers ce panorama, nous montrons également que notre sujet d'étude ne se laisse pas situer simplement et que selon les objectifs visés et les techniques mises en œuvre, l'inscription dans un domaine spécifique n'est pas acquise.

1.1 Panorama des applications informatiques traitant du langage naturel

Le schéma 1.1 propose une synthèse des systèmes actuellement disponibles pour le traitement automatique de l'information (schéma inspiré de Nazarenko (2005)).

L'intérêt de situer sur un même axe des systèmes aussi différents que ceux produits pour la recherche d'information ou pour l'extraction d'information est de mettre en avant leurs points communs (rechercher de l'information, utiliser des techniques proches) tout en montrant en quoi ils divergent (question de la thématique *vs* information précise et localisée).

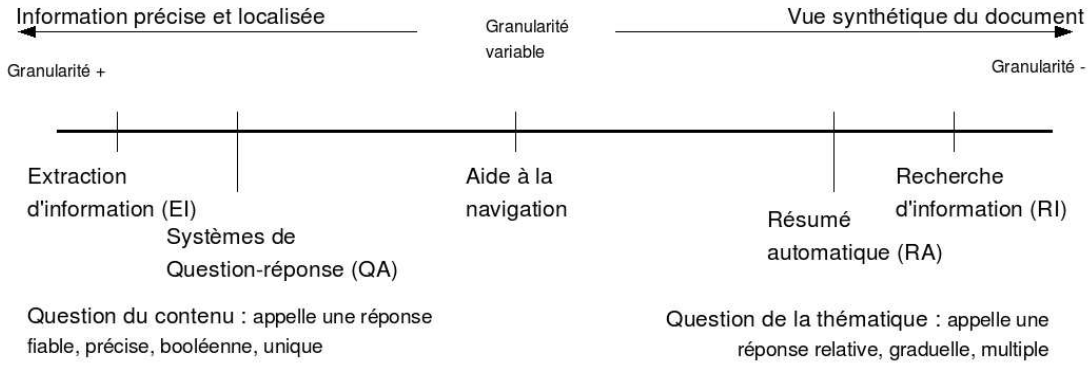


FIG. 1.1 - *Synthèse des systèmes de traitement de l'information*

À gauche de l'axe, l'extraction d'information se caractérise par la recherche d'une information précise et localisée. À l'autre extrémité de l'axe, la recherche d'une information propose une vue synthétique du document. Ainsi, selon l'objectif visé, les systèmes de traitement des connaissances viseront une granularité fine, variable ou à très large spectre.

Nous présentons enfin les systèmes d'aide à la navigation textuelle. Ces systèmes mettent en œuvre des techniques permettant notamment de traiter les textes selon des niveaux d'analyse variables. La variabilité des niveaux d'analyse est un point important dans notre travail.

1.1.1 La question de la thématique des documents : RI et RA

Recherche d'Information (RI)

La Recherche d'Information (ou RI) a pour objectif la sélection automatique des textes jugés pertinents pour une requête donnée. Le processus se fait en deux étapes : d'abord une indexation des documents puis un appariement entre les termes de la requête et ceux de la base.

Les descripteurs utiles à l'indexation sont extraits automatiquement des textes. La plupart du temps (et après suppression des mots grammaticaux ou/et des mots-outils), c'est l'ensemble des termes apparaissant dans un document qui serviront de descripteur. Ils sont généralement transformés (lemmatisation, racinisation, troncation). À l'aide de ce jeu de descripteurs, il est alors possible de représenter le document par un vecteur dans l'espace des termes.

Une fois les documents transformés, il faut rechercher ceux qui répondent le mieux à la question d'un utilisateur. Il existe plusieurs approches :

– L'*approche ensembliste* considère que l'ensemble des documents s'obtient par une série d'opérations (intersection, union et passage au complémen-

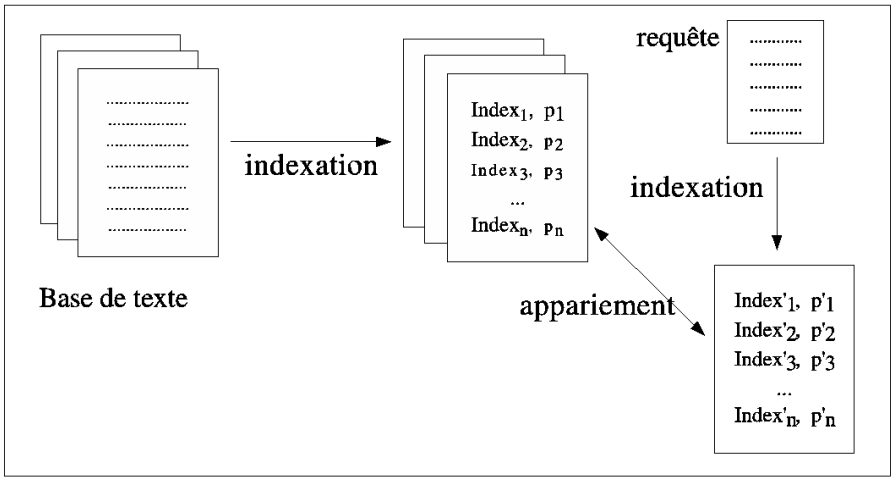


FIG. 1.2 - *Modèle général d'un système de RI*

taire).

- L'*approche algébrique* (ou vectorielle) considère que les documents et les questions font partie d'un même espace vectoriel. Le père de cette approche est Salton (Salton *et al.*, 1975) : son hypothèse est que les fréquences d'occurrences des termes d'un texte donnent une bonne représentation du contenu du texte. Les documents les plus pertinents sont ceux qui sont les plus proches de la requête.
- L'*approche probabiliste* cherche à modéliser la notion de pertinence[1] en considérant que les documents ne sont pas plus ou moins pertinents, mais que c'est leur probabilité de pertinence est plus ou moins importante (Jones *et al.*, 2000).

Il existe également des modèles capables d'interagir avec l'utilisateur dans le but d'améliorer progressivement les réponses du système de RI au cours d'une même session. L'utilisateur indique à chaque fois les documents pertinents pour sa question. Ces indications peuvent également servir à améliorer globalement le fonctionnement du système de RI.

Les conférences TREC, ou *Text REtrieval Conference*, sont des conférences annuelles d'origine américaine. Elles fournissent un cadre commun pour l'évaluation des systèmes automatisés de recherche d'information traitant en texte intégral de très volumineux corpus. D'année en année, cette conférence affine la méthode qu'elle a proposée et étend les mesures d'évaluation et de comparaison qu'effectuent depuis 1992 des participants issus du monde de la recherche et de celui

[1]Le taux de pertinence est le pourcentage exprimant le rapport entre le nombre de documents pertinents extraits et le nombre total de documents extraits. La pertinence se définit par rapport à l'évaluation d'un juge humain.

de l'industrie. Ces travaux sont essentiellement destinés aux concepteurs de ces systèmes d'accès à l'information textuelle (plus qu'à leurs utilisateurs), et visent à leur fournir des pistes de développement.

Au niveau de l'évaluation des outils, la RI a introduit deux mesures qui restent aujourd'hui des mesures incontournables lorsqu'on développe un outil en T.A.L. Il s'agit des mesures de précision et de rappel.

$$precision \ : \frac{P}{(NP+P)}$$
$$rappel \ : \frac{P}{(P+R)}$$

où NP est le nombre d'unités non pertinentes fournies par le système,
P est le nombre d'unité pertinentes
et R le nombre d'unités pertinentes dans la base et non fournies par le système.

Ces mesures permettent d'évaluer l'efficacité d'un document par rapport à une requête et de comparer les différents résultats renvoyés par les systèmes de RI. Un système de RI sera très précis si presque tous les documents renvoyés sont pertinents. Le rappel sera élevé s'il renvoie la plupart des documents pertinents du corpus pour une question. En général, plus un système de RI est précis, moins il a de rappel et inversement.

Plus généralement, ces mesures sont utilisées dans l'ensemble des systèmes de T.A.L. pour mesurer l'efficacité d'un système : nous nous en servirons également pour évaluer notre programme de repérage des expressions linguistiques (outil ALIDIS) puis pour évaluer le prototype final de recherche des segments d'obsolescence.

Résumé Automatique

Un résumé automatique de texte est une version condensée d'un document textuel obtenue de façon automatique.

Il existe deux approches principales pour générer des résumés de texte. L'approche *par abstraction* vise à rédiger un résumé en générant des phrases qui ne sont pas forcément contenues dans l'original. Cette approche est également nommée *approche par compréhension* car on cherche en amont à représenter le contenu informationnel. Cette approche est peu développée.

L'approche *par extraction* consiste à extraire des phrases complètes considérées comme les plus importantes d'un point de vue informationnel puis à les concaténer de façon à produire un extrait. Cette dernière approche est de loin celle qui est la plus utilisée dans les systèmes réels.

> *« Ce qui est appelé résumé automatique consiste en fait plutôt en un écrémage automatique. Il s'agit en effet d'attribuer à chaque phrase*

> *un « poids informationnel » en fonction de marques de surface : marques discursives (indications de plan,...), place dans le paragraphe, poids individuel des mots de la phrase en fonction de leur répartition dans tout le document... On construit alors le résumé par extraction des phrases de plus fort poids, dans l'ordre du document, et en respectant un seuil de réduction paramétrable. [...] par ailleurs les indices lexicaux et discursifs peuvent partiellement différer d'un domaine à l'autre. Enfin, ce qui est pertinent dans un texte dépend de l'utilisateur et doit donc être paramétrable en fonction de ce dernier. »* (Habert, 2005, p. 53)

Les niveaux mobilisés en termes de traitement linguistique proposent *a minima* un comptage fréquentiel des mots.

Dans les systèmes plus sophistiqués, la structure du texte et, plus précisément, l'endroit où apparaissent les éléments et les mots importants, sont pris en compte. Les positions d'introduction ou de conclusion sont souvent privilégiées (Mani, 2001; Marcu, 2000). Le degré de relation entre les phrases, la prise en compte de facteurs internes tels que la présence de phrases-types, de tournures particulières sont également des solutions exploitées. Cette prise en compte des éléments de structuration textuelle nous intéresse particulièrement.

Les traitements syntaxiques sont généralement assez rares. Dans de nombreux systèmes qui existent à l'heure actuelle, les phrases ne sont pas analysées syntaxiquement. La phrase est exploitée telle qu'elle apparaît dans le texte source et est considérée *a priori* comme syntaxiquement correcte (Mani, 2001). Ce parti pris entraîne souvent des difficultés quant au *lissage* de texte résumé pour le rendre cohérent et cohésif puisqu'il s'agit généralement de phrases saillantes (*i.e.* importantes) concaténées.

Les systèmes récents mettent de plus en plus en avant la nécessité de considérer les besoins des utilisateurs des systèmes, besoins variables d'un individu à l'autre, d'une situation à l'autre. Les systèmes de résumé automatique cherchent alors à prendre en compte les attentes des utilisateurs et proposent des versions adaptables. L'évolution des systèmes de résumé automatique vers des systèmes de navigation intra-documentaire est en cours : l'utilisateur est orienté, guidé dans la lecture des documents, relativement à une thématique donnée, un objectif particulier. Ceci permet de donner de la souplesse et une meilleure qualité dans la recherche des informations et en même temps d'être plus proche de la demande des utilisateurs. Jean-Luc Minel parle de *parcours de lecture*[2] (Minel, 2002; Couto et Minel, 2004).

L'évaluation des systèmes de résumé automatique est une tâche difficile pour laquelle la communauté a des réponses partielles. En effet, une évaluation automatique demande de disposer d'un système capable de générer des résumés de qualité humaine. Des solutions pragmatiques peuvent être envisagées comme le proposent

[2]Jean-Luc Minel sur `http://www.technolangue.net/article.php3?id_article=329`

les conférences NIST dont l'objectif est d'évaluer des systèmes de RA. La métrique ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) y a été développée : elle mesure la *couverture* entre les *N-grammes* produits automatiquement par une machine à ceux contenus dans des résumés écrits par un certain nombre de juges humains. Un haut niveau en ROUGE implique empiriquement un niveau de corrélation avec les résumés humains. Se pose dans ce contexte le problème du référentiel humain qui s'avère souvent variable selon les juges, leurs objectifs, leur façon d'envisager et de concevoir un résumé. Dans le chapitre 2, nous verrons que la question du jugement humain est également incontournable pour aborder le phénomène de l'obsolescence.

1.1.2 La question du contenu précis et local : EI et QR

L'Extraction d'Information (EI)

Nous allons maintenant décrire les systèmes d'extraction d'information que le schéma 1.1 (p. 16) situe à l'opposé des systèmes de recherche d'information.

Les systèmes d'EI ont pour objectif d'extraire des éléments de contenu spécifiques et prédéfinis, représentés sous une forme structurée afin d'intégrer des bases de données.

> *« The goal of information extraction (IE) is to build systems that find and link relevant information from natural language text ignoring irrelevant information. The information of interest is typically pre-specified in form of uninstantiated frames like structures also called templates. The templates are domain and task specific. The major task of an IE system is then the identification of the relevant parts of the text which are used to fill a template's slot. »* (site *Language Technology World*[3], Information extraction)

Une source d'informations linéaire et non structurée est transformée en une base de données structurée où seules les données pertinentes pour la tâche sont archivées. Dans des *templates*, sont rangées des informations diverses telles que le type du phénomène recherché, la date où il a eu lieu, l'endroit où cela s'est passé, qui a été mis en cause, quelles sont les relations entre les acteurs, etc.

Le niveau de connaissances linguistiques nécessaire pour mener à bien cette tâche est généralement plus élevé que dans le cas de la recherche d'information qui envisage, comme nous l'avons présenté précédemment, les textes comme des sacs de mots et génère une structure sous forme d'une liste plate d'index sans lien syntaxique ni sémantique.

Le repérage des entités nommées y est central : il s'agit de rechercher les noms propres de personnes, de lieux ou encore d'organisations, les expressions temporelles (dates, durées, horaires) et les noms de quantité (monnaies, mesures, pourcentages). Les systèmes plus élaborés cherchent à typer les entités nommées

[3] `http://www.lt-world.org/`

en valeurs : villes, fleuves, noms de personnalités (acteurs, présidents, etc.). Ces unités présentent la caractéristique d'être faiblement polysémiques et sont de fait des indices fiables. Leur fonctionnement est assez proche de ceux des noms propres et ils sont fortement liés à l'actualité. Leur repérage se fait soit à l'aide de dictionnaires, soit à l'aide de règles d'extraction contextuelles.

Il est important de noter que le traitement des entités nommées est devenu tellement central qu'il est souvent effectué au détriment du repérage d'expressions de point de vue, d'expressions de la modalité et d'expressions de l'argumentation (Nazarenko, 2005). Le repérage des entités nommées est également abordé dans notre recherche.

Les systèmes d'extraction d'information font appel soit à des méthodes statistiques soit à des méthodes linguistiques (Bouhafs, 2005). Nous nous intéressons aux méthodes exploitant des connaissances linguistiques.

Les méthodes linguistiques suivent généralement le modèle suivant : d'abord une analyse lexicale et morpho-syntaxique permet le découpage des textes en mots, leur lemmatisation et de leur étiquetage morphologique ; suit une analyse syntaxique qui établit les relations entre les mots, les syntagmes, les relations entre les syntagmes ; enfin, les actants et les phénomènes (entités nommées principalement) sont repérés et typés (relations) ; les techniques visant à *comprendre* les calculs de la référence, les inférences, les anaphores, sont alors utilisées.

La méthode la plus utilisée en EI est le *pattern-matching* : il s'agit de faire correspondre une valeur avec un motif associé. Par exemple, le motif textuel « x a acheté n actions de y » permettra de retrouver dans un texte les valeurs de x, n et y.

Il existe de plus en plus d'expériences cherchant à exploiter des indices génériques comme les structures énumératives ou plus largement la structure du document (Toussaint, 2004).

Les conférences MUC (*Message Understanding Conference*) ont pour objectif l'évaluation des systèmes d'extraction d'information pour lesquels des techniques robustes d'identification et de catégorisation des entités nommées sont primordiales (Daille et Morin, 2000).

Les systèmes de Question-Réponse (QR, *Question-Answering (QA)*)

Les systèmes de question-réponse (QR) renvoient, à partir d'une requête formulée en langage naturel sous forme de question, une réponse, la plus exacte possible. Comme pour l'extraction d'information, il ne s'agit pas de rechercher un document entier, mais de repérer des passages textuels courts, précis et pertinents pour une requête.

Les techniques exploitées dans les systèmes de Question-Réponse sont proches de celles de l'extraction d'information. Deux étapes sont nécessaires : tout d'abord l'analyse de la question (qui est généralement en langage naturel) puis la recherche et la formulation de la réponse.

L'analyse de la question se fait généralement par une recherche des mots-clés

pertinents de la question. Il faut également être capable de repérer les relations entre ces mots, ce qui passe souvent par une analyse syntaxique. Un repérage des entités nommées est ensuite mené dans l'ensemble des documents, comme on l'a vu pour l'extraction d'information. L'appariement, la mise en relation entre les mots-clés de la question et les termes extraits des documents est l'étape centrale des systèmes de Question-Réponse.

Le type de réponse retourné par le système peut alors être simplement l'extraction d'une phrase jugée pertinente ou, dans des systèmes plus sophistiqués la génération d'une réponse appropriée n'existant pas dans les documents.

> *« Answer extraction (AE) aims at retrieving those exact passages of a document that directly answer a given user question. AE is more ambitious than information retrieval and information extraction in that the retrieval results are short phrases, not entire documents, and in that the queries may be arbitrarily specific. It is less ambitious than full-fledged question answering in that the answers are not generated from a knowledge base but looked up in the text of documents. »* (site *Language Technology World*[4], Answer extraction)

Ce sont aussi les conférences TREC (*cf.* p. 17) qui ont porté ce type de problématique.

1.1.3 Une granularité variable : les systèmes d'aide à la navigation

Que l'on parte des systèmes de recherche d'information ou, à l'opposé, des systèmes d'extraction d'information, la tendance actuelle consiste en une conception granulaire, variable des documents.

Ainsi, que ce soit du côté des systèmes de question-réponse ou du côté du résumé automatique, la tendance est aujourd'hui à la navigation textuelle variable selon la demande des utilisateurs : le système doit être adaptable et revoit alors des élément textuels plus ou moins locaux, plus ou moins globaux. Ce sont les *parcours de lecture* proposés par Minel (2002) ou encore la navigation intra-documentaire telle que l'ont définie Enjalbert (2005, p. 353) et Bilhaut (2004).

L'idée de ce type de systèmes est de rendre compte de l'organisation des textes. Deux grandes orientations peuvent être mises au jour actuellement sachant que dans les deux cas, il s'agit de découvrir la structure interne des textes.

Le premier type de travaux sur l'organisation des textes peut être illustré par les travaux de Teufel (1999) et Teufel et Moens (2002) sur l'organisation argumentative des textes de type scientifique. L'auteur propose la notion d'*argumentative zoning* qu'elle définit comme la représentation de la structure rhétorique générale des textes pour orienter un traitement automatique d'EI ou de RA. Elle repère et annote sept types de zones à travers la détection d'indices textuels :

– *aim* : présentation de l'objet de l'article.
– *textual* : explicitation de la structure de l'article ou de la section.

[4] `http://www.lt-world.org/`

– *own* : description (neutre) de son propre travail relatif à l'objet d'étude (méthodologie, résultats, discussion).
– *background* : état de l'art des modèles et courants auxquels réfère l'étude.
– *contrast* : comparaison de l'étude avec d'autres travaux.
– *basis* : présentation des travaux sur lesquel l'étude se fonde.
– *other* : présentation neutre d'autres travaux.

Le repérage et l'annotation de ces zones permet alors d'envisager un parcours des textes scientifiques selon les besoins d'un utilisateur. Par exemple, l'utilisateur peut chercher à connaître les différents buts d'un article ou les critiques qu'il apporte ou encore rechercher la zone textuelle correspondant à une requête précise.

Cette méthode est basée sur le repérage et l'exploitation d'indices formels que nous décrivons dans la section 1.3.3). Elle utilise également un corpus segmenté et annoté manuellement par des juges humains. *In fine*, une phase d'apprentissage automatique des indices est menée à partir des données.

Le second type de travaux cherchant à rendre compte de l'organisation textuelle consiste en la recherche de *segments thématiques*. L'idée est de placer des marques aux endroits où un texte passe d'un thème à un autre. La segmentation thématique facilite l'extraction d'information : plutôt que de renvoyer le texte tout entier, seule la sous-partie correspondant au thème précis de la requête est mise en évidence.

> *« La répartition des segments obtenus en grands ensembles caractérisés chacun par un « air de famille » et par son manque de ressemblance avec les autres permet d'obtenir des thématiques que l'on peut chercher à hiérarchiser et à interpréter. Un thème peut donc se réaliser par des segments de texte discontinus. Le repérage des frontières peut s'appuyer sur un inventaire de marques linguistiques introductrices d'un nouveau thème ou sur la détection de rupture de cohésion lexicale. La segmentation thématique fournit une image de la structure sous-jacente d'un texte. »* (Habert, 2005, p. 52)

La segmentation thématique peut être entièrement issue de méthodes quantitatives ou numériques. Elle est dans ce cas basée sur la notion de cohésion lexicale (Halliday et Hasan, 1976) : la fréquence et la répétition des mots sont des indicateurs de l'homogénéité thématique du segment. Le *text-tiling* est la technique la plus connue pour ce type de recherches (Hearst, 1994)[5].

Un autre type de méthode, cette fois orientée linguistique, consiste en la prise en compte de marqueurs, d'indices linguistiques, et de formes typo-dispositionnelles qui vont être à même de fournir des informations sur la structure du document. L'objectif est toujours de fournir des méthodes permettant de passer facilement d'un niveau de granularité à un autre. Dans ce contexte de recherche, le projet REGAL a pour ambition de produire des résumés automatiques de façon dynamique en fonction des attentes du lecteur (Couto *et al.*, 2005; Hernandez et Grau, 2003;

[5]Dans Laignelet et Pimm (2007), nous mettons en place un outil utilisant le text-tiling afin de mesurer les apports d'une telle méthode pour l'obsolescence. Les résultats sont intéressants mais la méthodologie que nous avons mise en place mérite d'être améliorée.

Hernandez, 2004). Techniquement, ce projet associe les méthodes de segmentation thématique et la prise en compte de certains marqueurs linguistiques. Un système de visualisation des textes permettant la variabilité des niveaux de granularité et d'analyse a également été mis en place.

Les travaux développés dans le cadre du projet GEOSEM traitant de documents géographiques vont également dans le sens d'une segmentation thématique des documents. L'objectif est de délimiter des segments textuels croisant des critères différents : un phénomène situé dans un lieu et localisé à un moment précis (par exemple Bilhaut (2006)). Les techniques mises en œuvre dans ces travaux exploitent la notion d'indices linguistiques variés faisant appel à des niveaux de traitement hétérogènes : des temps verbaux (indice local) aux cadres de discours (indice à gros grain) en passant par les expressions temporelles ou encore la structure informationnelle.

Pour notre part, nous exploitons également l'idée qu'une expression linguistique, qu'un ensemble d'éléments textuels sont à même de guider le lecteur dans une interprétation orientée vers un but spécifique. Le texte est composé d'indices de natures différentes, à des niveaux de granularité variables qui structurent et organisent le texte. Il convient alors de repérer, d'identifier ceux qui sont pertinents pour un objectif particulier. La linguistique fournit alors des techniques intéressantes pour le développement de ce type de systèmes de T.A.L.

1.2 Techniques et méthodes en T.A.L.

1.2.1 Traitements de base

Dans tous les cas, des systèmes de RI à ceux d'EI, une analyse complète des documents est rarement menée. Sont souvent traités des « bouts de textes », des segments. Il s'agit d'analyses locales qui rendent difficile la prise en compte de phénomènes sémantiques à distance, comme la portée d'un adverbial temporel (Bilhaut, 2006). L'accent est souvent mis sur les aspects lexicologiques. D'une manière générale, les techniques utilisées dans les systèmes de traitement de l'information font appel aux types d'analyse suivants :

- *a* - étiquetage morpho-syntaxique : repérage les mots inconnus et identification des catégories des mots ;
- *b* - identification des entités nommées et typage des mots inconnus ;
- *c* - annotation syntaxique de surface : identification des termes et des relations de dépendance ;
- *d* - analyse de la structure du document (titres, rôles des différentes sections) ;
- *e* - calcul, mesures statistiques (redondance et hétérogénéité lexicale) ;
- *f* - repérage des structures d'emphase (importance de certains éléments textuels).

Parmi ces analyses, celles que nous mettons en œuvre exploitent les niveaux *a*, *b*, *d*, *e*, *f* (*cf.* partie III, p. 123) : nous procédons à un étiquetage morpho-syntaxique des textes puis au repérage d'entités nommées ainsi qu'à leur annotation sémantique ; nous exploitons également la structure des documents ainsi que certains

éléments textuels spécifiques ; enfin, nous utilisons des techniques statistiques. Le niveau *c* consistant en une analyse syntaxique des phrases n'est pas mis en place dans ce travail. Au vu de notre objectif de recherche d'indices et de combinaisons d'indices, il ne nous semble pas justifié dans un premier temps. Mais il n'est pas exclu qu'il faille à terme mettre en place une telle analyse. Ce type de traitement pourrait par exemple permettre le traitement des chaînes de référence ou celui des relations anaphoriques et ainsi apporter des informations discursives supplémentaires et probablement pertinentes.

Les niveaux d'analyse textuelle tendent à se diversifier et l'on commence à sortir d'une vision où le lexique prédomine au détriment des analyses syntaxiques qui considère souvent ces deux niveaux comme deux domaines séparés. On observe l'émergence d'analyses de type sémantique (dans les systèmes de navigation intra-documentaire notamment) qui associent les connaissances morphologiques, lexicales et syntaxiques.

Avant de présenter les travaux faisant intervenir des notions de sémantique, nous tenons à présenter les travaux de Biber (1988, 1989) car ils introduisent une façon originale (du moins au moment où ces travaux sont sortis) d'utiliser des outils linguistiques dans un objectif de classification automatique de textes (domaine généralement rattaché à la RI).

1.2.2 La démarche classificatoire émergente de Biber

La méthode de Biber (1988, 1989) est basée sur l'hypothèse qu'un texte porte en lui des collocations de traits qui permettent de le caractériser. Cette démarche classificatoire est une étude systématique de la variation linguistique. Elle est dite *émergente* car elle ne cherche pas à valider des types prédéfinis. Au contraire, l'identification des types émerge du traitement statistique de la caractérisation linguistique des textes. Ainsi, pour Biber, les types de textes sont définis comme des corrélations de caractéristiques linguistiques participant à une même fonction globale.

Le travail de Biber a porté sur les cooccurrences entre 67 traits linguistiques dans les 1000 premiers mots de 481 textes d'anglais contemporain écrits et oraux. Ces textes relèvent de genres divers : articles de recherche, reportages, conversations, nouvelles radiophoniques, etc. Les traits étudiés sont identifiés automatiquement et renvoient à seize catégories distinctes : marqueurs de temps et d'aspect, adverbes et locutions adverbiales de temps et de lieu, pronoms et proverbes, questions, passifs, modaux, coordinations, négations, etc. (Biber, 1988, p. 73).

La *statistique multidimensionnelle* est mise à contribution pour repérer les oppositions majeures entre associations de traits linguistiques : elle rassemble les traits qui ont tendance à apparaître ensemble et constitue dans le même temps les configurations de traits qui sont systématiquement évités par les mêmes rassemblements. Cette démarche permet d'obtenir des pôles multiples, positifs et négatifs correspondant à des constellations. Chaque texte, par son emploi des traits linguistiques étudiés, se situe en un point déterminé d'un espace à *n* dimensions. C'est en

production impliquée *vs* **production à visée informative** :
traits caractéristiques du premier pôle :
do comme pro-verbe, 1ère et 2ème personne,
be comme verbe principal, présent,
démonstratifs, contraction.
orientation narrative *vs* **non-narrative** :
traits caractéristiques du premier pôle :
passé, 3ème personne, participe présent,
verbes dicendi.
référence explicite *vs* **dépendante de la situation** :
traits caractéristiques du premier pôle :
propositions relatives objet et sujet,
coordinations phrastiques, nominalisations.
visée persuasive apparente *vs* **non apparente** :
traits caractéristiques du premier pôle :
infinitifs, modaux (prédiction, nécessité, possibilité),
verbes de persuasion, subordination conditionnelle.
information abstraite *vs* **non abstraite** :
traits caractéristiques du premier pôle :
connecteurs, passif, subordonnées réduites,
propositions circonstancielles.

TAB. 1.1 - *Traits linguistiques et dimensions*

ce sens que la démarche suivie est inductive et non déductive : les traits pertinents qui permettent d'opposer ou de rapprocher différents textes sont issus des textes et non d'un savoir qui leur serait extérieur.

La première étape de la démarche de Biber consiste dans le regroupement de traits linguistiques fréquemment en cooccurence dans les textes (Biber, 1989). Il met en évidence les cinq dimensions exposées dans le tableau 1.1 (p. 26).

La seconde étape utilise des techniques de classification automatique permettant de rapprocher les textes en fonction de leurs coordonnées sur les cinq dimensions. Huit types de textes sont dégagés par Biber sachant que, d'une part, les textes appartenant à chaque type doivent avoir en commun le maximum de caractéristiques linguistiques et que, d'autre part, les différents types doivent être le plus distincts possibles (*cf.* tableau 1.2, p. 27).

Selon Biber, les types de textes correspondent à des corrélations effectives entre traits linguistiques. Ils ne se confondent ni avec les typologies fonctionnelles ni avec les genres : ces derniers sont des catégories intuitives utilisées par les locuteurs pour répartir les productions langagières (Biber, 1989).

- **interaction interpersonnelle intime** (*" intimate interpersonal interaction »*) ;
- **interaction informationnelle** (*" informational interaction »*) ;
- **exposé scientifique** (*" scientific exposition »*) ;
- **exposé savant** (*" learned exposition »*) ;
- **fiction narrative** (*" imaginative fiction »*) ;
- **récit** (*" general narrative fiction »*) ;
- **reportage situé** (*" situated reportage »*) ;
- **argumentation impliquée** (*" involved persuasion »*).

TAB. 1.2 - *Types de textes*

Le travail de Biber a été mené sur des textes en anglais. Malrieu et Rastier (2001) proposent une adaptation de cette méthodologie pour des textes en français.

Notre travail ne cherche pas à fournir une catégorisation en termes de types de textes de notre corpus. Mais la démarche classificatoire de Biber est une démarche intéressante car elle envisage les formes linguistiques en termes de complémentarité, de configurations d'indices pour rendre compte d'un typage des textes spécifique : c'est également ce vers quoi nous tendons.

De plus, nous défendons l'intérêt de mener une méthodologie de type émergente pour traiter la question de l'obsolescence dans les textes : l'obsolescence est un phénomène encore vague et nous souhaitons apprendre des textes. C'est à partir des annotations manuelles et automatiques et des traitements statistiques que les indices et combinaisons d'indices vont émerger. Les indices et combinaisons d'indices pertinents ne sont pas définis *a priori*.

1.2.3 La Méthode d'Exploration Contextuelle

La Méthode d'Exploration Contextuelle (MEC), développée au LaLIC par Desclès (1996), cherche à répondre aux besoins du filtrage sémantique de textes en identifiant, indépendamment d'un domaine particulier, certaines informations sémantiques. Les auteurs cherchent à rendre ce modèle entièrement autonome et adaptable selon les besoins des utilisateurs (BenHazez *et al.*, 2001).

Ce modèle se base d'un côté sur l'identification dans les textes de marqueurs linguistiques d'une catégorie grammaticale ou discursive, et d'autre part sur une exploration du contexte des marqueurs identifiés. Cette méthode s'apparente aux techniques de *pattern-matching*.

La méthode d'exploration contextuelle s'inspire de différents travaux en linguistique sur la polysémie qui montrent que plusieurs valeurs sémantiques peu-

vent être attribuées à des morphèmes de l'imparfait ou du passé composé (Desclés *et al.*, 1991). Pour lever les ambiguïtés polysémiques, cette méthode attribue une valeur sémantique à une entité linguistique examinée (indicateur) en fonction de la présence dans son contexte d'autres entités linguistiques (indices). L'exploration contextuelle exploite uniquement le contexte de l'entité linguistique examinée sans analyse morpho-syntaxique préalable (Goujon, 2000).

Une règle d'exploration contextuelle est de la forme suivante :

Partie Condition :
Soit un marqueur pivot X
SI le contexte gauche de X est G
ET SI le contexte droite de X est D
Partie Action :
ALORS prendre la décision Y (fin ou non fin d'un segment)

Une règle d'exploration contextuelle est constituée de deux parties : une partie condition et une partie action ou décision. Les conditions s'appliquent en deux temps : dans un premier temps la règle repère un indicateur spécifique (unités linguistiques) à un problème traité ; dans un deuxième temps la règle recherche dans le contexte gauche et/ou droit les indices linguistiques éventuels qui vont permettre le déclenchement des actions. Une action consiste à attribuer une valeur sémantique à l'indicateur ou à la phrase le contenant.

Dans Desclès et Guentcheva (2003), la méthode d'exploration contextuelle est utilisée pour la résolution sémantique des significations du passé composé. Les auteurs y décrivent tout d'abord le réseau des signification du passé composé puis un extrait de règles contextuelles. La règle R1 par exemple a pour but le filtrage de la veleur d'événement pour un verbe au passé composé. Elle est de la forme :

Partie Condition :
Soit un marqueur pivot [*verbe au passé composé*]
SI le contexte gauche ou droit contient un adverbial comme { *à l'aube du jour*, *à* + article + {*jour de/ année de/ début/ fin/ milieu/ premier/ dernier* } + substantif, *autrefois*, *dès que*, *en même temps que*, *par moment*, *naguère*, *puis*, *un moment après*, ... }
Partie Action :
ALORS [attribuer la valeur *événement* au verbe]

Cette méthode permet *(i)* d'interpréter le contexte d'un marqueur linguistique, *(ii)* d'analyser la position d'un marqueur dans le texte (début de phrase, premier

paragraphe, etc.), *(iii)* de manipuler les éléments structurels du texte (titres, paragraphes, etc.) et enfin *(iv)* d'identifier la structure thématique.

Aujourd'hui, la méthode d'exploration contextuelle postule qu'il est possible de repérer certaines informations sémantiques à partir de marques de surface en réponse à des besoins spécifiques d'utilisateurs cherchant à sélectionner des informations importantes, comme par exemple repérer les actions dans des textes techniques (Garcia, 1998), repérer les relations causales entre des situations (Jackiewicz, 1998) ou encore identifier les définitions des termes proposés explicitement ou implicitement par un auteur ainsi que les annonces thématiques (Cartier, 1998).

Cette méthode est utilisée pour de nombreuses tâches :

- identification des valeurs aspectuelles d'une proposition : système SECAT (Desclès *et al.*, 1992) ;
- filtrage automatique de phrases importantes dans un texte en vue d'un résumé automatique : système SERAPHIN/SAPHIR (Berri *et al.*, 1996) ;
- modélisation des connaissances par analyse des marqueurs linguistiques de relation entre concepts : systèmes SEEK et SEEK JAVA (LePriol, 1999) ;
- acquisition de connaissances causales à partir de textes : système COATIS (Garcia, 1998) ;
- extraction de citation : système CitaRE (Mourad, 2000) ;
- segmentation de textes en phrases (Mourad, 2002).

La plate-forme FilText (BenHazez *et al.*, 2001) met en œuvre la méthode d'exploration contextuelle : elle vise à accueillir des connaissances linguistiques. FilText a ensuite été implémenté dans dans la plate-forme logicielle ContextO (Minel, 2002) qui regroupe ainsi les ressources linguistiques issues de systèmes antérieurs (Cartier, 1998; Garcia, 1998; Jackiewicz, 1998; Desclès *et al.*, 1992) soit environ 11 500 marqueurs et 250 règles d'exploration contextuelle qui sont intégrées progressivement (LePriol, 2000).

Cette méthode distingue le travail des linguistes de celui des informaticiens. Les linguistes sont chargés de construire les ressources et les règles et les informaticiens implémentent les travaux des linguistes. Par ailleurs, s'il est évident que les ressources linguistiques disponibles dans la plateforme ContextO sont importantes et linguistiquement utiles, elle ne sont malheureusement pas transposables d'une technologie à une autre. C'est pourquoi nous n'avons pas réutilisé ces ressources développées au LaLIC[6].

Pour notre part, nous exploitons la plateforme LINGUASTREAM[7] qui permet au linguiste-informaticien d'être autonome depuis la création des ressources jusqu'à leur formalisation dans des langages informatiques à portée (lexiques sémantiques, expressions régulières, macro-expressions régulières, ProLog en ce qui nous concerne). Des plateformes telles que GATE ou encore la plate-forme d'annotation ALVIS développée au LIPN[8] vont également dans le sens d'une modularité des

[6] Laboratoire Langues, Logiques, Informatique, Cognition

[7] *cf.* chapitre 5.1, p. 127.

[8] Laboratoire d'Informatique Paris-Nord

traitements (exploitation et intégration d'outils de T.A.L. et de terminologie, de ressources et de corpus).

La réutilisation des ressources existantes n'est pas toujours facile d'un côté parce que les ressources sont généralement créées pour et relativement à un domaine spécifique et d'un autre côté parce que les différentes technologies utilisée ne sont pas toujours réellement compatibles. D'où la nécessité d'utiliser des formats ouverts, libres de droit et disponibles[9].

1.3 La notion d'indice linguistique en T.A.L.

L'utilisation d'indices linguistiques est aujourd'hui largement exploitée dans les systèmes de T.A.L. Nous avons vu dans les sections précédentes comment les techniques et outils de T.A.L. favorisent des analyses linguistiques précises, plus ou moins locales, interdépendantes et permettant au final un accès localisé et précis dans les textes (entre autres, *cf.* le résumé automatique, p. 18, la navigation textuelle, p. 22 ou encore la Méthode d'Exploration contextuelle, p. 27).

Mais les techniques évoluent et l'exploitation d'indices se précise. Nous présentons dans cette section une série de travaux qui utilisent des indices de types et de natures différentes dont la particularité est de considérer la dimension phrastique et la dimension discursive de manière interdépendante.

1.3.1 Des analyses de type sémantique

Les travaux de Bilhaut (2006) s'inscrivent dans le domaine du traitement automatique des langues et concernent l'analyse sémantique de la structure du discours. Plus spécifiquement, l'accent est porté vers la question de l'analyse thématique : l'auteur envisage l'étude de la structure des textes selon des critères relatifs à la répartition du contenu informationnel dans les textes.

La notion de *thème*, notion à la fois complexe[10] et rarement considérée en tant qu'objet d'étude dans le domaine de la recherche d'information, est au centre des travaux de Bilhaut (2006). Il propose une approche originale du *thème* comme objet discursif, sémantique et structuré. Il définit le concept de *thème composite* comme un *« objet constitué de deux éléments : le premier, [appelé] noyau thématique ou simplement noyau, correspond au topique discursif du segment en tant que représentant de son à propos ; le second est lui-même un ensemble d'éléments [appelle] satellites et qui définissent l'univers de discours au sein duquel se tient le noyau. »* Les *axes sémantiques* correspondent à des espaces notionnels susceptibles de participer à l'indexation de l'information dans les textes considérés. Ils peuvent être génériques comme le temps ou espace, ou plus spécifiques à un domaine ou à une pratique (axe des niveaux scolaires, des types de transports, etc.).

[9] L'ensemble de ce travail est disponible à l'adresse suivante : `http://marion.laignelet.free.fr`, sous licence LGPL.

[10] L'auteur dresse un bilan des notions de *thème*, de *topique*, de *sujet* ou encore d'*à propos*, tant en linguistique qu'en sciences de l'information ou en traitement des langues.

L'auteur évalue son modèle à travers un cas d'étude spécifique sur le traitement sémantique des documents géographiques et plus précisément sur l'analyse automatique des cadres de discours spatio-temporels. Parallèlement à ces recherches, l'auteur a mis en place la plateforme LINGUASTREAM qui constitue un environnement d'expérimentation intégré permettant l'élaboration de modèles linguistiques opérationnels et la mise en œuvre de principes méthodologiques originaux. Nous utilisons LINGUASTREAM car elle permet notamment de mener des analyses linguistiques de types différents et à granularité variable.

1.3.2 Des analyses à granularité variable

Le travail de Widlöcher (2008) porte sur l'analyse des structures rhétoriques du discours, c'est-à-dire des *« stéréotypes organisationnels qui participent au cheminement argumentatif des textes »*. L'auteur vise la constitution d'un cadre théorique et opérationnel général, permettant la modélisation et l'exploration computationnelle de telles structures.

Il propose notamment d'articuler son analyse autour des trois notions que sont unités, relations et schémas, notions qui présentent des propriétés particulières : variabilité du grain, flexibilité, non-linéarité et non-séquentialité potentielles, interactions local/global, etc.

L'auteur propose également le formalisme CDML[11] qui permet de modéliser des structures discursives par l'expression de contraintes sur des objets textuels de différentes natures (morphologique, syntaxique, sémantique, etc.), à différents niveaux de grain. Un analyseur permet de projeter ces contraintes sur corpus pour identifier les structures décrites. Deux études de cas ont par ailleurs été entreprises, sur deux types de structures significativement différentes : la première porte sur l'hypothèse de l'encadrement du discours de Charolles (1997), et la seconde explore les relations de contraste à différentes échelles, entre des objets linguistiques variés.

Nous apprécions dans ce travail la pertinence de la réflexion sur les problèmes de la variabilité du grain d'analyse, de la non-linéarité et de la non-séquentialité potentielles des unités et de leurs relations, des interactions entre le niveau local et le niveau global. Ces problématiques, également au centre de notre travail, posent la question de leur théorisation et de leur mise en application concrète dans des systèmes informatiques.

1.3.3 Combinaisons d'indices : du phrastique au discursif

Les travaux présentés dans cette section ont comme point commun de mettre en œuvre des techniques d'apprentissage automatique sur des corpus annotés manuellement et automatiquement.

[11] *Constraint-based Discourse Modeling Language*

Les travaux de Teufel et Moens (2002). Nous avons déjà parlé des travaux de Teufel (1998, 1999) sur la détection d'indices textuels pour le repérage automatique de zones argumentatives (*cf.* section 1.1.3, p. 22).

Dans Teufel et Moens (2002), les auteurs vont plus loin en mettant au point un système d'apprentissage automatique de la structure argumentative de textes scientifiques. Leur méthode se base sur :

- un nombre important d'indices variés :
  * position de la phrase,
  * structure textuelle explicite (structure de la section, du paragraphe, du titre),
  * longueur de la phrase,
  * reprise des éléments du titre (mots du titre, mesure TF*IDF),
  * morphologie des verbes (voix, temps, mode),
  * présence de citation,
  * *history* (catégorie précédente la plus probable),
  * marqueurs méta-discursifs (*formulaic expressions*/*cue-phrases*, agent, action).
- un corpus d'apprentissage annoté par des juges humains selon les sept catégories rhétoriques suivantes : *aim*, *textual*, *basis*, *background*, *contrast*, *own*, *other* (*cf.* section 1.1.3).

L'objectif final de ce travail vise le résumé automatique de textes scientifiques.

Les travaux de Bouffier (2008). Ils sont guidés par un cadre applicatif précis, la modélisation des *Guides de Bonnes Pratiques Médicales*. L'auteure propose des outils informatiques facilitant l'accès aux connaissances contenues dans ces guides.

L'auteure cherche à montrer l'apport d'une approche textuelle qui s'affranchit de la simple prise en compte d'analyses locales des textes et exploite la structure du texte. Plus précisément, elle recherche les combinaisons d'indices pertinentes pour signaler l'existence ou non d'une relation conditionnelle entre deux segments. La stratégie proposée est fondée sur l'exploitation de connaissances linguistiques obtenues par une méthode liant observation linguistique et apprentissage artificiel.

L'auteure développe une application, *GemFrame*, qui automatise le processus de modélisation en fournissant une première représentation structurée de ces textes. L'application vise à extraire les segments exprimant une recommandation et ceux exprimant une condition puis à associer à chaque segment conditionnel l'ensemble des segments « recommandation » qui dépendent de cette condition. Les relations de dépendance s'établissent souvent au delà du niveau de la phrase, ce qui justifie le recours à une approche textuelle.

Le système a été validé sur trois aspects complémentaires : utilité, performances et pertinence de la méthode.

Les travaux de Lucas et Crémilleux (2003). L'objectif de ces travaux est la détection automatique de fautes. Les auteurs exploitent des marqueurs linguistiques caractérisés par leur niveau de hiérarchie textuelle (relation entre marques linguistiques et mise en forme matérielle. Le même type de descripteurs est utilisé dans Zerida *et al.* (2006) où l'objectif est de caractériser trois types de textes biomédicaux (articles de recherche, de synthèse, cliniques). Les auteurs observent une différence significative dans l'organisation de l'écrit et dans le style des trois types de textes biomédicaux.

Dans Zerida *et al.* (2006), deux notions fondamentales sont mises en évidence :

- la *notion de position* : les mots n'ont pas le même rôle ni la même importance suivant leur place dans le document (titre, résumé, introduction, etc.) ; leur importance varie également suivant leur position dans la partie, le paragraphe, la section, etc.
- la notion d'*héritage du contexte* : chaque segment hérite des descripteurs positionnels et formels qui caractérisent la mesure englobante. Ainsi, une phrase *connaît* ses descripteurs, ceux du paragraphe, ceux de la partie, etc.

Dans ce travail. Notre système respecte et utilise de façon centrale cette notion d'héritage du contexte (*cf.* section 6.1.3, p. 166).

La méthodologie que nous mettons en place est en quelques points similaire à celles de Teufel et Moens (2002) et Bouffier (2008) (*cf.* partie III, p. 123) : annotation manuelle des segments recherchés, observation linguistique pour l'exploitation d'indices pertinents, repérage et annotation automatique des indices sémantiques et discursifs et enfin apprentissage automatique sur les données.

La principale différence avec ces travaux est que nous cherchons à caractériser des phénomènes différents. En effet, les segments d'obsolescence, contrairement aux segments de recommandation et de condition ou aux segments argumentatifs, présentent la particularité de ne pas être des objets linguistiques.

De plus, les indices linguistiques, discursifs et structurels que nous exploitons sont plus nombreux et plus diversifiés. À l'image des travaux de Lucas et Crémilleux (2003); Zerida *et al.* (2006), nous exploitons notamment des indices structurels et positionnels.

Concernant la phase d'apprentissage automatique, notre utilisation du classifieur consiste, non pas à supprimer des descripteurs qu'un système jugerait peu pertinents mais à découvrir des combinaisons d'indices sur la base de tous les descripteurs réellement présents dans le corpus d'apprentissage. De plus, nous ne faisons aucun *a priori* sur les combinaisons de descripteurs possibles : nous laissons le système les *révéler* du corpus de manière automatique.

Enfin, notre modèle de traitement des données textuelles est entièrement reproductible et les outils développés sont disponibles et peuvent être réutilisables pour d'autres tâches.

1.4 Conclusion

Ce chapitre a été l'occasion de situer l'application visée - le repérage de zones contenant de l'information potentiellement obsolescente - dans le vaste champ des systèmes de traitement de l'information. Nous nous rapprochons des techniques d'aide à la navigation intra-documentaire : la granularité des segments d'obsolescence est variable ; ils peuvent être très locaux (une date, un chiffre) ou plus globaux (de la taille de la phrase voire du paragraphe). Nous ne cherchons pas à extraire une information précise comme cherchent à le faire les systèmes d'Extraction d'Information ni à rendre compte de la thématique d'un texte comme le fait la Recherche d'Information. Nous cherchons à repérer des segments présentant une caractéristique particulière, celle de contenir une information potentiellement obsolescente.

Nous faisons l'hypothèse que les segments d'obsolescence sont signalés dans les textes par des indices linguistiques précis (Laignelet, 2007), de niveaux variés et de types différents (Laignelet, 2006b,a) et qui fonctionnent en combinaisons. La plateforme LINGUASTREAM permet d'appréhender et d'articuler ces indices linguistiques et discursifs hétérogènes. Mais c'est l'apparition et l'exploitation de ces indices en configuration qui nous intéresse particulièrement. Et pour répondre à cette problématique, nous avons mis en place un système d'apprentissage automatique de règles linguistiques et discursives pour le repérage de segments obsolescents.

Dans le chapitre suivant, nous présentons le corpus d'apprentissage ainsi que la phase d'annotation manuelle qui a été entreprise sur ces données textuelles. Ce chapitre 2 permet d'ouvrir la réflexion sur la question de l'obsolescence et de sa caractérisation en termes linguistiques : le chapitre 3 (p. 65) propose une description fine et détaillée des divers éléments linguistiques qui semblent pertinents pour caractériser l'obsolescence et le chapitre 4 (p. 89) présente les modèles et hypothèses linguistiques sur lesquels nous nous sommes fondés pour mener notre réflexion sur l'obsolescence.

Chapitre 2

Délimitation du phénomène de l'obsolescence sur la base d'un corpus annoté

Pour nous donner les moyens de décrire finement le phénomène de l'obsolescence tel que nous le définissons, la *linguistique de corpus* propose une méthodologie adaptée : c'est à partir de textes réels et d'annotations d'experts que nous avons mené une étude exploratoire et descriptive sur la notion d'obsolescence.

> *« Aucun corpus ne représente la langue : ni la langue fonctionnelle qui fait l'objet de la description linguistique, ni la langue historique qui comprend l'ensemble des documents disponibles dans une langue. En revanche un corpus est adéquat ou non à une tâche en fonction de laquelle on peut déterminer les critères de sa représentativité et de son homogénéité. »* (Rastier, 2005)

Il est important d'insister sur le fait que le corpus mis en œuvre n'a pas été constitué dans le but de servir une analyse linguistique *a priori*. Le critère de regroupement des textes est exclusivement fonctionnel : ils sont tous de type encyclopédique et une partie des informations contenues devront un jour ou l'autre potentiellement subir une mise à jour ; ils sont par ailleurs issus de documents réels, en l'occurrence des encyclopédies publiées sur le marché de l'édition francophone.

2.1 Constitution du corpus

Les textes sur lesquels nous travaillons ont été collectés parmi le fonds éditorial de deux maisons d'édition : les Éditions ATLAS et les Éditions LAROUSSE.

Leur sélection a d'abord été conditionnée par le matériel textuel que les maisons d'édition ont mis à disposition : toutes les encyclopédies ne sont pas forcément dans un format informatique exploitable et/ou toutes les ressources ne peuvent pas être mises à disposition pour des raisons éditoriales et/ou politiques.

Parmi les textes disponibles, la sélection a ensuite été faite selon les deux critères suivants :

- les textes devaient être de type encyclopédiques ;
- ils devaient présenter un développement relativement long (*a minima* une dizaine de phrases et un titre) pour permettre notamment la prise en compte d'indices *à gros grain*.

Les textes réunis sont des documents réels, issus d'une situation authentique et spécifique (l'édition encyclopédique) : ce sont ces mêmes textes qu'il faut/faudrait/faudra mettre à jour pour une publication future.

> *« All the material included in a corpus, whether spoken, written or gathered along any intermediate dimensions (Biber 1988) is assumed to be taken from genuine communications of people going about their normal business. »* Tognini-Bonelli (2001, p.55)

Le corpus élaboré dans le cadre de ce travail n'a donc pour objet que la description des textes de type encyclopédique et plus précisément de la notion d'obsolescence et non la langue dans sa globalité.

Les fiches [ATLAS] réunies sont au nombre de soixante-dix, ce qui représente une faible proportion du nombre de fiches encyclopédiques existant réellement à la vente (2400 fiches). Ce *faible* nombre de fiches est lié à deux causes : premièrement, il a été difficile de réunir les CDs contenant les fiches ; deuxièmement, le travail de normalisation des données a été long et fastidieux (*cf.* p.37).

En ce qui concerne le fonds Larousse, trois dictionnaires encyclopédiques ont été mis à notre disposition : le *Petit Larousse Illustré*, le *Grand Universel Larousse* et le *Grand Larousse Illustré*. Nous ne travaillons que sur le *Grand Universel Larousse* et le *Grand Larousse Illustré* car ils sont les seuls à contenir des développements encyclopédiques longs.

La question de la représentativité du corpus mérite d'être soulevée : le corpus élaboré est-il représentatif du type encyclopédique ? Les connaissances que nous découvrirons au fil de ce travail seront-elles transposables à d'autres textes issus d'autres encyclopédies ?

> *« A corpus is representative when the findings based on its contents can be generalized to a larger hypothetical corpus. »* (Leech, 1991 in Tognini-Bonelli (2001, p.57))

Si pour des raisons techniques mais aussi temporelles, il n'est pas possible d'exploiter d'autres textes dans l'immédiat, il sera nécessaire de valider ce travail sur de nouvelles données. Nous pensons en particulier à utiliser le fonds encyclopédique disponible à travers l'encyclopédie en ligne WIKIPEDIA. Dans cette perspective, la démarche proposée vise la reproductibilité des traitements (*cf.* chapitre 6, p. 161).

Le fait que les textes soient issus de deux sources différentes (Atlas et Larousse) garantit une certaine fiabilité des analyses ainsi que de la représentativité du corpus pour le type encyclopédique.

De plus, la taille du corpus est conséquente (10 000 phrases, soit 282 000 mots[1]) si l'on considère d'un côté qu'il s'agit de données réelles issues du monde professionnel et de l'autre que nous mènerons des analyses discursives et sémantiques complexes sur ces données (*cf.* chapitre 5, p. 125).

2.2 Description du corpus [ENCYCLO]

Le corpus dans son ensemble est nommé « corpus [ENCYCLO] » ; il est composé de deux sous-corpus, [ATLAS] et [LAROUSSE].

2.2.1 Prétraitement et normalisation des corpus

S'agissant de textes issus du monde professionnel, une phase de prétraitement des données, de normalisation des formats est nécessaire. D'une manière générale, les textes du corpus [ENCYCLO] sont transformé en XML à l'aide du langage de transformation XSLT. Le format XML présente l'avantage d'être aujourd'hui un standard.

De plus, parce qu'il permet de séparer le contenu de la présentation et de se concentrer sur la structure, ce format nous a permis de traiter de manière uniforme des objets très divers car issus de deux entreprises différentes.

Enfin, il est le format requis dans la plateforme LINGUASTREAM utilisée dans la cadre de ce travail (*cf.* section 5.1, p. 127).

Le sous-corpus [ATLAS]

Concernant le corpus ATLAS, nous sommes partis du format initial Quark XpressMac.

Le logiciel Quark XPress présente l'avantage de posséder une option d'aide à la transformation XML en se basant sur une DTD[2] et les définitions des styles. Cette aide est loin d'être négligeable même s'il a fallu contrôler minutieusement le résultat de chacune des fiches transformées. Ce contrôle manuel, fiche par fiche, consiste à vérifier qu'il ne manque pas de portions de textes et que la mise en page est correctement conservée. Nous faisons le choix de conserver un maximum d'informations de typo-disposition (titres, paragraphes, gras, italiques, texte encadré, surlignements, etc.) car nous exploitons ces informations dans les traitements automatiques ultérieurs.

Comme l'illustre l'exemple de la figure 2.1, la typo-disposition des fiches est importante. Ces éléments ont en eux-mêmes une valeur informative qu'il est important de mettre en évidence : c'est notamment le cas des encadrés à droite de la

[1] À titre de comparaison, le corpus mis en œuvre dans les travaux de Bouffier (2008) comporte 25 Guides de Bonne Pratique soit environ 150 000 mots.

[2] La Document Type Definition (DTD), ou Définition de Type de Document, est un document permettant de décrire un modèle de document SGML ou XML.

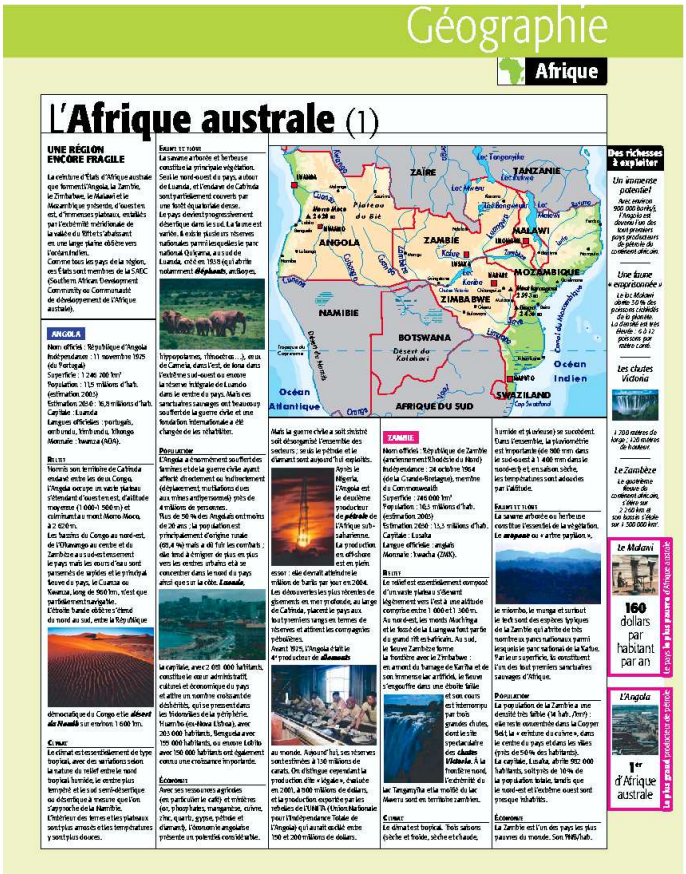
fiche[3], des titres colorés et encadrés ou encore des images.

La figure 2.2 montre un extrait de fiche en XML après sa transformation. Les informations typo-dispositionnelles sont conservées : par exemple, les titres colorés et encadrés sont dans des balises $< titreCadre > ... < /titreCadre >$, le texte des encadrés à droite de la fiche sont dans les balises $< colonne > ... < /colonne >$, et là où une image est insérée se trouvent les balises $< legendeSchema > ... < /legendeSchema >$.

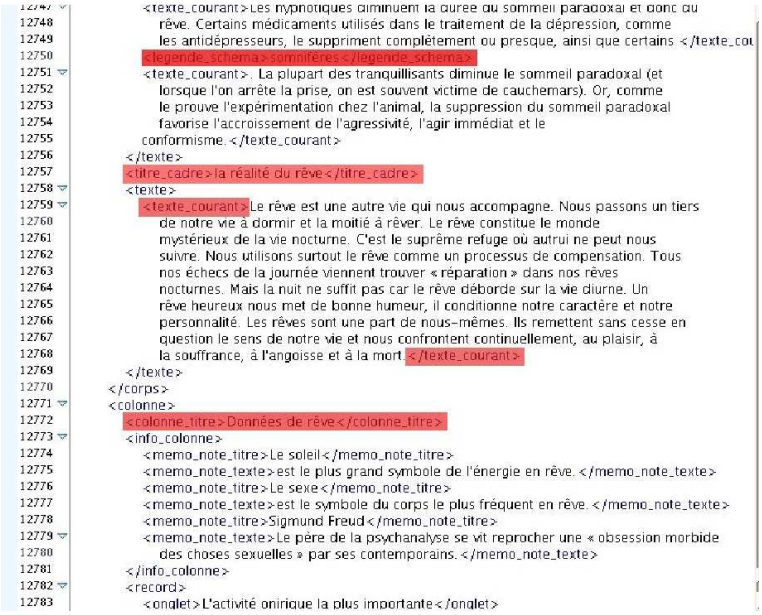
Le sous-corpus [LAROUSSE] ([GLI] et [GUL])

Le sous-corpus [LAROUSSE] est constitué de deux encyclopédies : le *Grand Larousse Informatisé* ([GLI]) et *le Grand Universel Larousse* ([GUL]).

[3]Il s'agit des « Mémo-Notes » : cette partie de la fiche constitue une sorte de zone d'appel du lecteur et a pour objectif éditorial de constituer un résumé de la fiche.



Exemple 2.1 - *Un exemple de fiche des Éditions Atlas*



Exemple 2.2 - *Un exemple de fiche des Éditions Atlas après transformation vers le format* XML

Il a été plus aisé de constituer ces sous-corpus puisque nous sommes partis des sources XML fournies par la société Larousse.

Nous avons ensuite procédé à des transformations XSLT pour ne garder que les articles des dictionnaires encyclopédiques qui contiennent des zones encyclopédiques de taille moyenne ou longue. La typo-disposition est relativement légère : nous gardons les informations concernant les titres et leur niveau hiérarchique. Les informations de type *Rubrique*[4] sont également conservées, comme c'est le cas pour les fiches du corpus ATLAS.

La figure 2.3 (p. 40) est un exemple de page de l'*Encyclopédie Universelle Larousse*. La figure 2.4 (p. 41) présente un extrait du sous-corpus [LAROUSSE] après transformation vers le format XML.

[4]La rubrique fait référence au domaine de l'entrée encyclopédique : il peut s'agir de géographie, d'histoire, de littérature, d'économie, de sciences, etc.

ESCRIMER

L'ESCURIAL PH. © R. FREDOCCIOMOS

ESOPE entouré des animaux de ses fables, par Francis Barlow (1626-1702). Gravure. Bibliothèque nationale de France, Paris. PH. © CELL. ARCHIVES LAROUSSE

« Le silence éternel de ces espaces infinis m'effraie. »

CITATION – Blaise Pascal, *Pensées*, 206. Cette parole, selon toute vraisemblance, est mise dans la bouche du libertin, avec lequel Pascal envisageait de dialoguer dans son ouvrage. Chaque citation des *Pensées* porte en référence un numéro. Celui-ci est le numéro que porte dans l'édition Brunschvicg – laquelle demeure aujourd'hui la plus généralement répandue – le fragment d'où la citation est tirée.

880

c'est-à-dire de pointe mais aussi de taille et de contre-taille. C'est une arme de convention : la surface valable est constituée de toute la partie du corps située au-dessus de la ligne horizontale passant par le sommet des hanches. La priorité est donnée à l'attaque : l'action offensive initiale exécutée en allongeant le bras et menaçant continuellement la surface valable de l'adversaire. L'arbitre du match est chargé de donner les commandements, de faire respecter le règlement relatif au protocole et au matériel, et d'attribuer les touches. Il peut sanctionner des fautes de combat à l'aide d'un carton jaune (simple avertissement), rouge (touche de pénalité) ou noir (exclusion immédiate). [COMPLÉTÉ INTERNET]

ESCRIMER (S') v.pr. (à). Faire tous ses efforts en vue d'un résultat difficile à atteindre ; s'appliquer, s'évertuer. *S'escrimer à faire des vers.*

ESCRIMEUR, EUSE n. Personne qui pratique l'escrime.

ESCRIVÁ DE BALAGUER (Mgr Josemaría ou José María), *Barbastro 1902 - Rome 1975*, prélat espagnol. Fondateur de l'Opus Dei (1928), il fut béatifié en 1992 et canonisé en 2002.

ESCROC [eskro] n.m. Personne qui escroque.

ESCROQUER v.t. (ital. *scroccare*, décrocher). S'emparer de qqch d'une façon malhonnête, frauduleuse. *Escroquer des millions.*

ESCROQUERIE n.f. **1.** Action d'escroquer. **2.** Délit consistant à s'approprier le bien d'autrui par des tromperies ou des manœuvres frauduleuses.

ESCUDERO (Vicente), *Valladolid 1892 - Barcelone 1980*, danseur et professeur espagnol. Il fut le partenaire de la Argentina, avec laquelle il créa *l'Amour sorcier* (1925).

ESCUDO [eskudo] n.m. **1.** Unité monétaire principale du Cap-Vert. **2.** Ancienne unité monétaire principale du Portugal. (Devenu, dès le 1er janvier 1999, une subdivision de l'euro, l'escudo portugais a cessé d'exister, au profit de la monnaie unique européenne, en 2002.)

ESCULAPE MYTH. ROM. Dieu de la Médecine. Il correspond à l'Asclépios grec.

Escurial (l'), en esp. **el Escorial,** palais et monastère d'Espagne, au pied de la sierra de Guadarrama, au N.-O. de Madrid.
◆ Accomplissement d'un vœu de Philippe II après sa prise de Saint-Quentin, conçu comme nécropole royale et centre d'études au service de la Contre-Réforme, il fut élevé de 1563 à 1584 par Juan Bautista de Toledo, l'Italien Giambattista Castello et Juan de Herrera dans un style classique sévère. Nombreuses œuvres d'art : bronzes des Leoni père et fils (Leone et Pompeo), peintures de primitifs flamands, de Titien, du Greco, de Ribera, de Velázquez, fresques de L. Giordano, tapisseries de Goya, notamment.
[RECHERCHE MULTIMÉDIA], [COMPLÉTÉ INTERNET]

ESDRAS ou **EZRA,** *Ve s. av. J.-C.*, prêtre juif. Il restaura la religion juive et le Temple après l'exil de Babylone.

ÉSÉRINE n.f. Alcaloïde végétal toxique, tiré du physostigma, utilisé naguère en médecine.

ESGOURDE n.f. *Arg.* Oreille.

ESHKOL (Levi), *Oratou, Ukraine, 1895 - Jérusalem 1969*, homme politique israélien. Il fut Premier ministre de 1963 à 1969.

ESKILSTUNA, v. de Suède, près du lac Mälaren ; 90 726 hab. Importante ville industrielle : aciers spéciaux, coutellerie. Instituts de technologie. — Musées.

ESKIMO adj. inv. et n.m. → ESQUIMAU.

ESKIMO → ESQUIMAUX.

ESKIŞEHIR, v. de Turquie, à l'O. d'Ankara ; 482 793 hab. Centre commercial, avec quelques industries (agroalimentaire et mécanique).

ESMEIN (Adhémar), *Touvérac, Charente, 1848 - Paris 1913*, juriste français. Remarquable historien du droit, il est l'auteur d'importants travaux sur le droit public et le droit canon.

Esméralda (la), personnage du roman de V. Hugo *Notre-Dame de Paris* (1831-1832), qui est une jeune et belle bohémienne.

ESNAULT-PELTERIE (Robert), *Paris 1881 - Nice 1957*, ingénieur français.
◆ Il se distingua dans le domaine de l'aviation comme constructeur de monoplans robustes et comme auteur de nombreuses inventions, parmi lesquelles le moteur en étoile à nombre impair de cylindres et le dispositif de commande appelé « manche à balai » (1906). Il fut aussi un précurseur de l'astronautique, théoricien et ardent propagandiste des voyages interplanétaires au moyen de fusées, allant jusqu'à prévoir l'utilisation de l'énergie nucléaire comme moyen de propulsion. (Acad. des sc., 1936.)
[COMPLÉTÉ INTERNET]

ESNÈH ou **ISNA,** v. d'Égypte (Haute-Égypte), sur le Nil. Vestiges de la salle hypostyle du temple ptolémaïque (colonnes gravées d'importants textes relatifs au mythe de la création).

ESO (European Southern Observatory), organisation européenne de recherches astronomiques dans l'hémisphère Sud, créée en 1962. Son siège est à Garching (Allemagne). Au Chili, l'ESO dispose d'un observatoire, sur le mont La Silla, et d'un très grand télescope (*VLT*), sur le Cerro Paranal.

ÉSOPE, *VIe s.-VIe s. av. J.-C.*, fabuliste grec. On attribue à ce personnage à demi légendaire un ensemble de *Fables*, connues des Athéniens dès la fin du Ve s. av. J.-C., qui inspirèrent Phèdre, les fabulistes du Moyen Âge et, plus tard, La Fontaine.

ÉSOTÉRIQUE adj. (*gr. esôterikos*, réservé aux seuls adeptes). **1.** Relatif à l'ésotérisme. CONTR. : *exotérique*. **2.** *Fig.* Peu compréhensible par le commun des mortels ; hermétique, obscur.

ÉSOTÉRISME n.m. **1.** Partie de certaines philosophies anciennes qui devait rester inconnue des non-initiés. **2.** Caractère ésotérique, obscur de qqch.

1. ESPACE n.m. (lat. *spatium*). **I.** Étendue indéfinie qui contient et entoure tous les objets. **2.** MATH. Ensemble de points, de vecteurs, etc., sur lequel on a défini une structure. *Espace vectoriel, topologique.* — *Spécial.* Espace vectoriel euclidien à trois dimensions. *Géométrie dans le plan, dans l'espace.* **3.** PSYCHOL. Représentation de cette étendue ; ce que nos sens en connaissent. *Espace auditif, visuel.* **4.** Volume occupé par qqch. *Ce meuble occupe peu d'espace.* **5.** Intervalle de temps. *Dans l'espace d'un an.* **6.** Surface, milieu affectés à une activité, à un usage particuliers. ◇ *Espace vert* : jardin, parc d'une agglomération. — *Espace publicitaire,* ou *espace* : portion de surface ou plage de temps destinée à recevoir de la publicité. *Achat, vente d'espace.* **7.** (Calque de l'all. *Lebensraum*). *Espace vital* : ensemble des territoires correspondant aux besoins d'expansion d'un État et devant lui être rattachés, selon une théorie nationaliste ; *fam.*, espace nécessaire pour ne pas se sentir gêné par les autres. **8. a.** Milieu dans lequel se meuvent les astres. **b.** Domaine situé au-delà de la partie de l'atmosphère terrestre où peuvent circuler les aéronefs ; ensemble des activités industrielles ou de recherche se rapportant à ce domaine. *Conquête de l'espace.* **c.** DR. *Espace aérien* : partie de l'atmosphère dont un État contrôle la circulation aérienne.

8 L'ESPACE ET SON EXPLORATION

L'exploration – certains disent la conquête – de l'espace a commencé le 4 oct. 1957, avec le lancement du premier satellite artificiel. Depuis ont été mis sur orbite plus de 4 000 satellites, parmi lesquels des vaisseaux occupés par des spationautes.

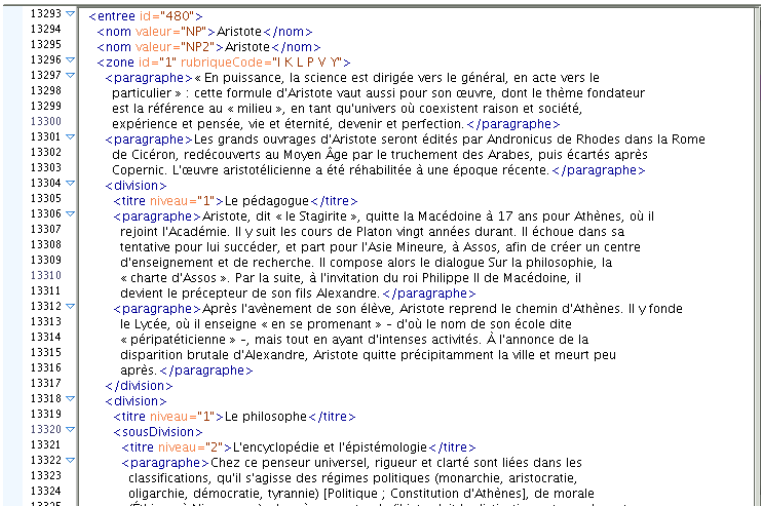
Un milieu hostile. L'espace est un milieu extrêmement hostile quique caractérise par l'absence d'atmosphère (*vide spatial*), des écarts de température considérables (selon qu'une surface est tournée vers le Soleil ou qu'elle se trouve dans l'obscurité, sa température peut passer de + 150 °C à – 150 °C) et la présence de rayonnements, de particules et de poussières microscopiques. Les satellites sont conçus pour résister à ces conditions hostiles ; l'homme ne peut survivre dans l'espace qu'en étant protégé par une enceinte étanche et pressurisée (vaisseau spatial et, pour les sorties extravéhiculaires, scaphandre). Tout engin spatial en mouvement orbital se trouve en état *d'apesanteur* (ou *impesanteur*), dont les conséquences sont spectaculaires dans les vaisseaux habités : les spationautes ne peuvent plus marcher, car leur poids n'est plus sensible, et les objets flottent autour d'eux s'ils ne sont pas fixés.

Les puissances spatiales. On appelle « puissance spatiale » un État ou un groupe d'États ayant prouvé sa capacité à placer des satellites en orbite à l'aide de ses propres lanceurs et qui peut conduire en toute autonomie l'activité spatiale de son choix. Chronologiquement, l'URSS a été la première puissance spatiale en plaçant sur orbite le premier satellite artificiel de la Terre, Spoutnik I, le 4 oct. 1957. Elle a précédé les États-Unis (1958), la France (1965), le Japon (1970), la Chine (1970), le Royaume-Uni (1971), l'Europe réunie au sein de l'Agence spatiale européenne (1979), l'Inde (1980) et Israël (1988). Le Royaume-Uni n'a procédé au lancement que d'un seul satellite à l'aide d'un lanceur de sa fabrication et a abandonné (en 1973) son programme de lanceur national. La France, elle aussi, a abandonné (en 1975) sa filière de lanceurs nationaux (Diamant), au profit du lanceur européen Ariane.

Astronautes et cosmonautes. Appelés *astronautes, cosmonautes* ou *taïkonautes*, selon qu'ils ont été passagers de vaisseaux spatiaux américains, russes ou chinois, ou encore *spationautes* selon la terminologie française, ils sont plus de 400 hommes ou femmes à avoir voié dans l'espacé. Leur sélection a partout été fondée sur les mêmes critères : une forme physique et psychique irréprochable associée à un solide bagage scientifique ou technique. La sélection est suivie d'une période d'entraînement comprenant un complément de formation théorique, des exercices physiques variés et des simulations en tous genres (d'accélération en centrifugeuse ; d'apesanteur à l'intérieur d'avions en vol balistique ; du « travail dans l'espace » en piscine, etc.).
Environ 10 % des sujets ayant volé dans l'espace ne sont ni des Américains ni des Russes ; ils appartiennent à une vingtaine de nationalités.

Les sorties extravéhiculaires. Le 18 mars 1965, A. Leonov, copilote du vaisseau spatial soviétique *Voskhod 2*, est devenu le premier « piéton de l'espace » : muni d'un scaphandre et relié au vaisseau spatial par un « cordon ombilical » de plusieurs mètres, il évolua pendant 12 min dans l'espace, satellisé à 28 000 km/h autour de la Terre. Depuis, les sorties dans l'espace se sont multipliées, atteignant couramment aujourd'hui des durées de

Exemple 2.3 - *Un exemple de page de l'Encyclopédie Universelle Larousse*



```
<entree id="480">
  <nom valeur="NP">Aristote</nom>
  <nom valeur="NP2">Aristote</nom>
  <zone id="1" rubriqueCode="I K L P V Y">
    <paragraphe>« En puissance, la science est dirigée vers le général, en acte vers le
    particulier » : cette formule d'Aristote vaut aussi pour son œuvre, dont le thème fondateur
    est la référence au « milieu », en tant qu'univers où coexistent raison et société,
    expérience et pensée, vie et éternité, devenir et perfection.</paragraphe>
    <paragraphe>Les grands ouvrages d'Aristote seront édités par Andronicus de Rhodes dans la Rome
    de Cicéron, redécouverts au Moyen Âge par le truchement des Arabes, puis écartés après
    Copernic. L'œuvre aristotélicienne a été réhabilitée à une époque récente.</paragraphe>
    <division>
      <titre niveau="1">Le pédagogue</titre>
      <paragraphe>Aristote, dit « le Stagirite », quitte la Macédoine à 17 ans pour Athènes, où il
      rejoint l'Académie. Il y suit les cours de Platon vingt années durant. Il échoue dans sa
      tentative pour lui succéder, et part pour l'Asie Mineure, à Assos, afin de créer un centre
      d'enseignement et de recherche. Il compose alors le dialogue Sur la philosophie, la
      « charte d'Assos ». Par la suite, à l'invitation du roi Philippe II de Macédoine, il
      devient le précepteur de son fils Alexandre.</paragraphe>
      <paragraphe>Après l'avènement de son élève, Aristote reprend le chemin d'Athènes. Il y fonde
      le Lycée, où il enseigne « en se promenant » - d'où le nom de son école dite
      « péripatéticienne » -, mais tout en ayant d'intenses activités. À l'annonce de la
      disparition brutale d'Alexandre, Aristote quitte précipitamment la ville et meurt peu
      après.</paragraphe>
    </division>
    <division>
      <titre niveau="1">Le philosophe</titre>
      <sousDivision>
        <titre niveau="2">L'encyclopédie et l'épistémologie</titre>
        <paragraphe>Chez ce penseur universel, rigueur et clarté sont liées dans les
        classifications, qu'il s'agisse des régimes politiques (monarchie, aristocratie,
        oligarchie, démocratie, tyrannie) [Politique ; Constitution d'Athènes], de morale
```

Exemple 2.4 - *Un exemple de fiche des Éditions Larousse (GLI) après transformation vers le format* XML

2.2.2 Taille du corpus [ENCYCLO]

Le corpus [ENCYCLO] constitue un fonds textuel de 282 200 mots, soit 10 000 phrases annotées manuellement selon leur caractère obsolescent ou non. Le tableau 2.1 (p. 42) décrit chacun des sous-corpus selon leur taille absolue et relative.

	[ATLAS]	[GLI]	[GUL]	[ENCYCLO]
Nombre de mots	193 606	26 237	62 298	282 141
Pourcentage des mots	68 %	10 %	22 %	100 %
Nombre de phrases	7 143	885	1 889	9 917
Pourcentage des phrases	72 %	9 %	19 %	100 %

TAB. 2.1 - *Taille du corpus*

2.3 Annotation manuelle du corpus

L'objectif de cette annotation manuelle est de rendre compte, à travers le regard des experts, de ce qu'est un segment d'obsolescence. L'intérêt de cette phase d'annotation manuelle est double : d'un côté, elle permet de créer un matériau pertinent pour la description linguistique de l'obsolescence (*cf.* chapitres 3 et 4) ; de l'autre côté, ce corpus est essentiel pour la phase d'apprentissage automatique des configurations d'indices en tant que base d'apprentissage supervisé (*cf.* chapitre 7).

2.3.1 Définition de la tâche d'annotation manuelle : repérer les segments d'obsolescence

Définir la tâche de mise à jour des informations est délicat. Cette tâche réelle et professionnelle est encore mal délimitée par les acteurs qui pourtant l'exécutent au quotidien dans leur travail.

D'une manière générale, il est relativement fréquent (que ce soit dans le monde de l'entreprise ou dans le monde de la recherche) de vouloir des outils qui automatisent une tâche précise alors que cette même tâche n'a été ni correctement décrite ni précisément délimitée. C'est la situation à laquelle nous avons dû faire face : même si la question de la mise à jour des encyclopédies constitue une problématique centrale pour nos deux partenaires, la nature même de cette tâche n'est pas clairement identifiée.

De plus, la description d'une tâche est parfois complexifiée par sa nature même. Suite aux entretiens passés avec les partenaires, il semble que la tâche de mise à jour des informations soit une activité intellectuelle caractérisée par une subjectivité relativement importante : elle semble par conséquent sujette à variation d'un individu à l'autre et parfois même pour un seul et même individu lorsqu'il est amené à reconsidérer une annotation effectuée auparavant.

Malgré ces difficultés, les nombreux entretiens passés avec nos partenaires ont permis de tirer des enseignements fructueux sur la question de la mise à jour et plus précisément sur la définition du phénomène de l'obsolescence. Avant de présenter le protocole d'annotation manuelle des segments obsolescents, il est important de revenir sur le phénomène de l'obsolescence en insistant notamment sur ce qui constitue sa nature même (les aspects évolutifs temporels) mais également sur ses limites et ses difficultés.

L'obsolescence, un phénomène complexe

La caractéristique incontournable pour définir le caractère obsolescent d'une information est sans aucun doute celle de la **relation au temps**.

La communication écrite implique nécessairement deux représentations pour un même événement : celle de la production du message et celle de sa réception. Le moment de l'écriture, *i.e.* le moment où le texte encyclopédique est rédigé et/ou publié, ne correspond donc pas au moment de la lecture de ce même texte.

Le moment de rédaction est généralement donné par la date de publication du document. Le moment de lecture est inconnu mais on le sait nécessairement postérieur au moment de rédaction. Idéalement, la distance entre ces deux moments doit être la plus courte possible comme l'indique le schéma 2.5 (p. 43). Une mise à jour effective des documents permet de déplacer sur l'axe du temps le moment de rédaction afin qu'il soit toujours le plus proche possible du moment de lecture. Alors que dans certains cas, plus cette distance est grande, plus la mise à jour se justifie (en géographie par exemple), dans d'autres cas, plus la distance est grande moins la mise à jour est nécessaire (en histoire notamment).

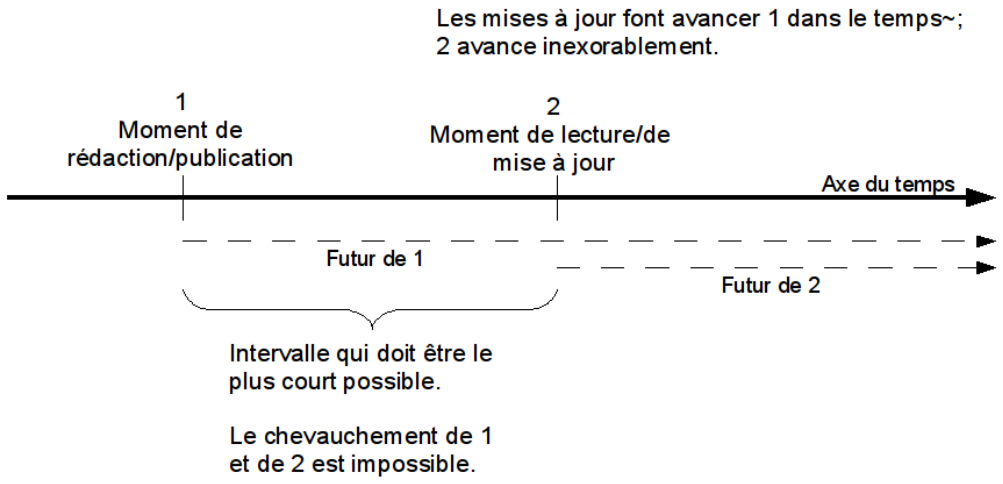


FIG. 2.5 - *La problématique temporelle de la mise à jour dans l'édition*

Le temps exprimé dans une encyclopédie est un temps indissociable de celui

de notre réalité. Les repères temporels peuvent être très éloignés, comme c'est le cas dans les articles historiques, ou, au contraire, très proches comme dans les articles géographiques ou de société. Il ne s'agit pas de textes *fictifs* mais de textes *encyclopédiques* qui relatent des événements de l'Histoire donc présumés réels. La réalité temporelle est bien la même dans les deux cas, une encyclopédie étant une sorte de photographie de la réalité (des connaissances) du monde à un moment donné. Dans ce type de textes, c'est l'information factuelle qui prédomine ; l'information peut être également envisagée sur le mode de l'irréel ou de l'hypothétique.

Les exemples (construits) suivants illustrent cette problématique :

(2.1) Aujourd'hui, le PIB par habitant de la France est de 27 600 dollars.

(2.2) En 2004, le PIB par habitant de la France est de 27 600 dollars.

Dans l'exemple 2.1, sachant qu'on est actuellement en 2009, et que le lecteur va naturellement interpréter l'adverbial déictique « aujourd'hui » comme étant l'année en cours[5], l'information est fausse puisque le PIB de la France le plus actuel (chiffres de 2008) est de 33 800 dollars.

À l'inverse, l'exemple 2.2 montre un cas où l'information restera toujours vraie : en 2004, le PIB par habitant de la France sera toujours de 27 600 dollars. Une mise à jour éditoriale sera cependant nécessaire si l'objectif du rédacteur est de fournir les résultats les plus récents par rapport à la date en cours : il faudra donc actualiser à la fois la référence temporelle (« En 2009 ») et la valeur du PIB.

Parallèlement aux aspects temporels, la question de la **validité de l'information** se pose lorsqu'on recherche l'obsolescence. Il s'agit de s'interroger non pas sur la qualité ou la vérité ou fausseté de l'information mais sur l'intérêt informationnel véhiculé par certaines phrases. Ainsi, dans l'exemple 2.6, l'association entre la date (« 2003 ») et le chiffre (« 67,7 millions d'habitants ») restera vrai, et ce peu importe le moment où l'information est lue. Mais ce qui est mis en avant par l'auteur, c'est la comparaison entre deux situations ayant lieu à des moments différents (« 2003 » et « 1927 »). Or, on est en droit de penser que ce qui intéresse le lecteur, ce sont les chiffres les plus actuels et non ceux de 2003. Mais alors, la comparaison sera-t-elle toujours pertinente ?

> En 2003, la population turque s'élève à **67,7 millions d'habitants**. Une forte poussée démographique a eu lieu au cours du xxe siècle : ils n'étaient que 13,6 millions **en 1927**. [...]
>
> Source : ATLAS

Exemple 2.6 - *La pertinence de l'information : le cas des comparaisons*

Pour notre part, même si le segment en question ne nécessite pas de mise à jour, il doit être malgré tout repéré afin d'être vérifié.

Lorsqu'on cherche à définir l'obsolescence, le **point de vue du rédacteur** constitue également une caractéristique importante. Il est fréquent que l'auteur ex-

[5] En l'occurrence, 2009.

prime ses doutes, ses espoirs quant au phénomène qu'il traite. Il peut également être interrogatif quant à l'évolution d'un phénomène.

Dans l'exemple 2.7, le rédacteur émet un jugement implicite dans sa description. Ici, il craint la désertification d'une région, il évoque des menaces qui pèsent sur la vie de la forêt. Est-ce pour autant à mettre à jour ? Ses craintes se sont-elles vérifiées ? Nous considérons ce type de segments comme des segments d'obsolescence.

> Le nord du pays **est guetté** par la désertification due à la fois à des facteurs d'ordre naturel et à des modes d'exploitation inappropriés (surpaturage) qui ne laissent pas le temps à la flore de se régénérer. Au sud, l'une des forêts tropicales humides les plus riches du continent (plus de 1 500 espèces végétales dans la réserve du Dja) **est menacée** de déforestation par la surexploitation des richesses en bois, très convoitées, qui représentent, après le pétrole, l'une des principales sources de revenu du pays. [...]
>
> Source : ATLAS

Exemple 2.7 - *Le point de vue du locuteur*

Ces trois caractéristiques, la question du temps, la question de la validité informative et la question de la subjectivité de l'auteur, forment le socle commun à la définition de l'obsolescence que nous avons proposé à nos partenaires. Elles servent de guide pour l'annotation manuelle du corpus. Mais savoir quoi annoter n'est pas suffisant : il est essentiel de déterminer le type, la taille du segment textuel minimal qui sera annoté.

La délimitation des segments textuels : syntagme, phrase, paragraphe, section ?

L'obsolescence est un phénomène qui est susceptible de se concrétiser tant au niveau d'une section entière que d'un paragraphe, d'une phrase ou encore d'un syntagme. Puisque nous visons un traitement automatisé des annotations manuelles, il est nécessaire de définir une unité d'analyse minimale : pour des raisons techniques notamment (*cf.* chapitre 6, p. 161), c'est l'unité phrase qui nous semble être la meilleure unité minimale pour l'annotation manuelle des segments d'obsolescence. En effet, même si une section entière ou un paragraphe est à mettre à jour, il est toujours possible de les diviser en phrases.

Comme l'illustre l'exemple 2.8, ce choix pose malgré tout quelques problèmes.

> Dans les premières années de la recherche, le risque de contamination était d'environ 25 %, il est aujourd'hui de moins de 2 %. [...]
>
> Source : ATLAS

Exemple 2.8 - *La délimitation des segments : phrase ou syntagme ?*

Dans cette phrase, la seconde partie uniquement est à mettre à jour (« il est

aujourd'hui de moins de 2 % »). Ce cas est assez fréquent lorsque le rédacteur compare entre deux états d'un même événement.

C'est parce que nous avons besoin d'un système d'annotation manuelle homogène que l'unité phrase a été retenue pour rendre compte de l'obsolescence.

La question des titres

La question des titres constitue un point important de notre travail. L'observation du corpus nous amène à considérer le cas des titres comme des éléments textuels particuliers. Alors qu'ils sont à même de contenir des indices linguistiques et discursifs propres à l'obsolescence, nous avons cependant constaté qu'ils ont très nettement un rôle de prédicteur de segments obsolescents et non une nature à être des segments obsolescents. Il est en effet relativement rare que le titre en lui-même soit obsolescent, et ce, même si toute la partie qu'il introduit est à mettre à jour.

Nous montrons dans le chapitre 4 (p. 89) que les titres sont des éléments importants pour le repérage de l'obsolescence mais qu'ils ne se situent pas au même niveau que les phrases[6]. Nous prenons le risque de passer à côté de titres obsolescents et préférons considérer ces objets textuels à part comme des prédicteurs de segments obsolescents. Nous revenons également sur ce point au chapitre 6 (p. 161).

2.3.2 Protocole d'annotation

Pour comprendre avec précision la tâche de mise à jour des informations dans les fiches encyclopédiques, un repérage manuel des zones textuelles a été mené. Les annotateurs sont :

- pour le sous-corpus [LAROUSSE], trois rédacteurs des Éditions Larousse et moi-même ;
- pour le sous-corpus [ATLAS], moi-même.

[6]Étant donné que nous recherchons des configurations d'indices, traiter les titres comme des objets différents des phrases donne la possibilité de faire hériter sur chaque phrase d'une section les caractéristiques du titre qui la régit.

Le protocole d'annotation est le suivant :

> Tâche :
> Annoter les segments textuels qui vous semblent nécessiter aujourd'hui ou à l'avenir une mise à jour de l'information.
> Contraintes :
> 1) Tout segment correspond à une phrase graphique, *i.e.* qui commence par une majuscule et se termine par un signe de ponctuation fort, et ce, même si le segment comprend également des informations ne nécessitant aucune mise à jour.
> 2) Un titre ne sera pas annoté.

Les consignes sont volontairement lâches et guidées par la tâche : en quelque sorte, il est demandé aux experts d'imaginer l'outil idéal d'aide à la mise à jour et ce qu'ils en attendent.

Concernant le sous-corpus [ATLAS], l'annotation des segments susceptibles de contenir de l'information évolutive a été effectuée par moi-même. Linguiste-informaticienne de formation, la principale difficulté que j'ai rencontrée a été de mettre de côté mes *a priori* sur la langue et sur la question de l'obsolescence. Il a au contraire été nécessaire de mettre en avant mon intuition sur la tâche de mise à jour proprement dite.

Concernant le corpus [LAROUSSE], nous avons fait appel aux connaissances et compétences d'experts au sein de l'entreprise Larousse. Ils nous ont permis de faire avancer la problématique de manière considérable. Ainsi, en plus de mon annotation, trois experts des Éditions Larousse ont effectué les annotations manuelles sur ce sous-corpus.

Les annotations ont toutes été menées à la main sur des versions papier des corpus et au surligneur. Elles ont ensuite été retranscrites au format XML.

Spécificités de l'annotation du corpus [ATLAS]

Sur l'ensemble des 69 fiches ATLAS, 47 fiches ont été annotées. Il reste 22 fiches qui constituent une partie du corpus de test destiné à l'évaluation finale du système (*cf.* chapitre 9, p. 217). Par domaines, le nombre de fiches annotées manuellement est :

- géographie : 10 fiches sur 14,
- médecine et santé : 11 fiches sur 13,
- sciences et techniques : 8 fiches sur 8,
- société : 6 fiches sur 7,
- sport : 3 fiches sur 5,
- histoire : 4 fiches sur 9,
- art et littérature : 3 fiches sur 8,

– faune et flore : 2 fiches sur 5.

Spécificités de l'annotation des corpus [GLI] et [GUL]

Pour ce qui est des sous-corpus [GLI] et [GUL], 42 entrées ont été annotées manuellement. Par domaines, le nombre de fiches annotées manuellement est :

– économie : 9 entrées,
– droit : 3 entrées,
– histoire : 11 entrées,
– géographie : 11 entrées,
– société : 11 entrées,
– faune et flore : 2 entrées,
– sciences et techniques : 5 entrées,
– médecine et santé : 4 entrées.

Une même entrée peut parfois se subdiviser en plusieurs rubriques : par exemple une entrée pour un pays a trois zones, Société, Géographie et Histoire. Ceci explique que les chiffres ci-dessus dépassent largement les 42 entrées annoncées.

La figure 2.9 montre un extrait du sous-corpus [LAROUSSE] dans lequel ont été reproduites les annotations manuelles des experts.

```
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet type="text/css" href="larousse.css"?>
<corpusApprentissage>
  <encyclopedie type="GLI">
    <entree id="232">
      <nom valeur="NC"> avortement </nom>
      <zone id="1" rubriqueName="Med">
        <paragraphe> On distingue l'avortement spontané (couramment appelé « fausse couche ») et
          l'avortement provoqué, pratiqué soit pour des raisons thérapeutiques, soit pour des
          raisons non thérapeutiques et appelé alors interruption volontaire de grossesse (IVG)
          lorsqu'il est légal. </paragraphe>
        <division>
          <titre niveau="2"> L'interruption volontaire de grossesse </titre>
          <paragraphe>
            <obsol id="209" annotateurId="13"> Dans de nombreux pays, l'interruption
              volontaire de grossesse est permise, si elle est pratiquée avant la 12e ou la
              14e semaine suivant l'arrêt des règles. </obsol> Elle est pratiquée sous
            anesthésie locale ou générale, par aspiration endo-utérine, avec une canule ou une
            seringue (méthode de Karman), ou, beaucoup plus rarement en raison des risques de
            lésions de la muqueuse utérine, par curetage. <obsol id="210" annotateurId="13">
              L'aspiration peut être remplacée jusqu'au 49e jour d'aménorrhée, en France, par
              un traitement associant le mifepristone (RU 486) et un dérivé des
              prostaglandines. </obsol>
            <obsol id="211" annotateurId="13"> Au-delà de 98 jours d'aménorrhée, une
              interruption volontaire de grossesse n'est plus autorisée dans la plupart des
              pays. </obsol>
          </paragraphe>
        </division>
        <division>
          <titre niveau="2"> L'avortement thérapeutique </titre>
          <paragraphe> Il se pratique à tout moment de la grossesse, sur demande des deux
            parents ou d'un seul, si la vie de la mère est en danger (insuffisances cardiaque,
```

FIG. 2.9 - *Annotation manuelle du sous-corpus* [LAROUSSE]

Les annotateurs experts ont rencontré des difficultés à respecter la consigne stipulant qu'il faut annoter des phrases entières et non des syntagmes ou des expressions locales. Ils ont également annoté des titres lorsque la section entière était à mettre à jour. Nous n'avons pas retranscrit l'annotation des titres dans les corpus

informatisés car cela représente surtout un moyen d'économie pour ne pas surligner une section dont la taille peut atteindre une cinquantaine de phrases.

Enfin, face à la difficulté de se représenter parfois le phénomène de l'obsolescence, ils ont parfois ajouté des commentaires du type « à vérifier si la situation politique le demande ». Ce type de commentaire est intéressant si par ailleurs les phrases potentiellement obsolescentes sont également annotées. Or ces remarques avaient la plupart du temps vocation à remplacer la délimitation des segments d'obsolescence. Dans ces cas-là, nous avons agi au cas par cas et la retranscription des annotations n'a été faite que si elle était justifiée par rapport à la tâche (et souvent au regard des annotations des autres experts).

L'obsolescence, un phénomène flou

Certaines annotations manuelles invitent à s'interroger sur la nature de l'obsolescence dans certains segments. Nous avons constaté des cas *limites*, des phrases où l'on est en droit de se demander si oui ou non, il est nécessaire de mettre à jour l'information. Par exemple, l'exemple suivant présente deux phrases annotées par un annotateur seulement dans lesquelles la présence d'une information potentiellement obsolescente n'est pas évidente.

(2.3) En 2001, le républicain G. W. Bush devient président. Le 11 sept., les États-Unis sont frappés par des attentats sans précédent, prenant pour cibles les tours jumelles du World Trade Center, à Manhattan, qui sont détruites, et le Pentagone, à Washington.

Dans l'exemple suivant, quatre annotateurs ont jugé que la phrase méritait d'être mise à jour. Or, il semble que ce choix est vraiment contestable et nous supposons que les événements politiques actuels ont orientés ces jugements.

(2.4) En dépit d'une situation économique et sociale difficile (déficits commercial et budgétaire record, augmentation de la pauvreté), G. W. Bush est réélu à la présidence en 2004.

Ces exemples montrent que l'obsolescence est un phénomène graduable et non consensuel. Ces constat sur les annotations manuelles nous ont amené à mettre en place une évaluation quantifiée de l'accord entre les juges-rédacteurs (*cf.* section suivante).

2.4 Premiers constats sur les segments d'obsolescence

Un des objectifs de l'annotation manuelle du corpus[7] est de permettre de mieux comprendre le phénomène de l'obsolescence, de le quantifier au sein des textes et de le décrire avec précision. Il s'agit donc en d'autres termes de permettre une validation de l'objet d'étude en tant qu'objet linguistique.

[7] L'autre objectif est d'utiliser ces annotations pour la phase d'apprentissage supervisé que nous développons au chapitre 7 (p. 175).

Après une évaluation quantitative de l'obsolescence, nous proposons une définition et une description de la notion de segment d'obsolescence.

2.4.1 Évaluation quantitative du phénomène de l'obsolescence

Proportion des segments d'obsolescence dans l'ensemble du corpus

Pour le corpus [LAROUSSE], cette phase d'annotation manuelle a mis en évidence la présence de 581 segments d'obsolescence dans les 42 fiches parcourues soit 20,2 % de phrases obsolescentes.

Pour le corpus [ATLAS], cette phase d'annotation manuelle a mis en évidence la présence de 927 segments d'obsolescence dans les 47 fiches parcourues soit 12.9 % de phrases obsolescentes.

	ATLAS	**LAROUSSE**	**ENCYCLO**
nombre total de phrases	7142	2874	9916
nombre de phrases obsolescentes	927	581	1508
pourcentage de phrases obsolescentes	12.9 %	20,2 %	15,2 %

TAB. 2.2 - *Proportion de segments obsolescents dans le corpus* ENCYCLO

Les chiffres de ce tableau montrent que le sous-corpus [LAROUSSE] ([GLI] et [GUL]) a une proportion plus élevée de segments obsolescents que le sous-corpus [ATLAS].

Ceci est dû exclusivement au fait que le sous-corpus [LAROUSSE] a bénéficié d'une multi-annotation et que ce tableau rend compte de tous les segments ayant été annoté au moins une fois par un annotateur. Si on considère le pourcentage de segments obsolescents par annotateur, cela donne les chiffres suivants, plus proche de la réalité des données :

- annotateur 1 : 268 segments obsolescents sur 2874 phrases, soit 9,3 %
- annotateur 2 : 274 segments obsolescents sur 2874 phrases, soit 9,5 %
- annotateur 3 : 173 segments obsolescents sur 2874 phrases, soit 6 %
- annotateur 4 : 382 segments obsolescents sur 2874 phrases, soit 13,3 %

Ces chiffres montrent que les annotateurs n'annotent pas de la même manière : les annotateurs 1 et 2 ont des comportements globaux assez semblables tandis que l'annotateur 3 a plutôt tendance à peu annoter et que l'annotateur 4 annote beaucoup.

Accord inter-juges au regard du sous-corpus [LAROUSSE]

Procéder à une annotation multi-juges nous permet de mesurer le degré d'accord entre les jugements d'obsolescence des experts.

Lorsqu'on cherche à mesurer l'accord entre des jugements, on considère que l'accord observé entre un ou plusieurs jugements qualitatifs résulte de la somme

d'une composante « aléatoire » (*i.e.* liée au hasard, P_{att}) et d'une composante d'accord « véritable » (ou réel, P_{obs}). D'où le calcul de la proportion d'accord observé (P_{obs}) et celui de la proportion d'accord aléatoire (P_{att}).

Le cœfficient Kappa (K) est traditionnellement utilisé dans ce cas. Il évalue la qualité ou l'intensité de l'accord réel entre deux juges (P_{obs}) tout en retirant la proportion de hasard ou de subjectivité de l'accord (P_{att}). C'est un indice qui permet de mesurer la reproductibilité des annotations, c'est-à-dire le degré avec lequel deux annotateurs différents vont pouvoir produire la même annotation[8].

L'accord observé (P_{obs}) correspond à la proportion des individus classés dans des classes diagonales concordantes. Il se fait sur la base d'une matrice de confusion dans laquelle sont indiqués le nombre de segments jugés obsolescents par les juges 1 et 2 et le nombre de segments jugés non obsolescents par les juges 1 et 2 (*cf.* tableaux 2.3 et 2.4).

		juge 1		
		obsol	non obsol	total
juge 2	obsol	A	B	X
	non obsol	C	D	Y
	total	X'	Y'	Z

TAB. 2.3 - *Une matrice de confusion pour le calcul de P_{obs}*

		juge 1		
		obsol	non obsol	total
juge 2	obsol	148	126	274
	non obsol	120	2397	2517
	total	268	2523	2791

TAB. 2.4 - *Accords observés entre les juges 1 et 2*

Les chiffres de ce tableau 2.4 montrent que dans 86 % des cas, les juges sont d'accord pour dire qu'un segment n'est pas obsolescent et dans 5 % des cas, ils évaluent de la même manière l'obsolescence d'un segment. En d'autres termes, dans 91 % des cas, les juges 1 et 2 sont d'accord. Ce fort pourcentage entraîne

[8] Il serait également intéressant de mesurer la stabilité des annotations en mesurant le degré avec lequel un même annotateur est capable de produire la même annotation à deux moments différents. Pour des raisons de temps, nous n'avons pas été en mesure de procéder à cette évaluation mais elle devra être mise en place.

nécessairement une forte valeur de P_{obs} comme nous allons le voir ci-après.

$$P_{obs} = \frac{A+D}{A+B+C+D}$$
Soit dans notre cas : $P_{obs} = 0.91$

Pour le Kappa (K), il nous faut maintenant calculer la proportion d'accords aléatoires, *i.e.* dûs au hasard (P_{att}). Nous nous basons sur les formules explicitées dans la matrice 2.5, qui, toujours sur la base des accord entre les juges 1 et 2, permet de construire la matrice 2.6.

		juge 1		
		obsol	non obsol	total
juge 2	obsol	$\frac{X*X'}{Z}$	$\frac{Y'*X}{Z}$	X
	non obsol	$\frac{X'*Y}{Z}$	$\frac{Y*Y'}{Z}$	Y
	total	X'	Y'	Z

TAB. 2.5 - *Une matrice de confusion pour le calcul de* $Patt$

		juge 1		
		obsol	non obsol	total
juge 2	obsol	**26**	248	274
	non obsol	242	**2275**	2517
	total	268	2523	**2791**

TAB. 2.6 - *Accord attendu entre les juges 1 et 2*

La concordance attendue P_{att} est alors égale à la somme des effectifs théoriques des deux classes concordantes, divisée par la taille de l'échantillon :

$$P_{att} = \frac{\frac{X*X'}{Z} + \frac{Y*Y'}{Z}}{Z}$$
Soit $P_{att} = 0.82$

Sur la base des calculs P_{obs} et P_{att}, on peut maintemant calculer le cœfficient

Kappa (K) :

$$K = \frac{P_{obs} - P_{att}}{1 - P_{att}}$$

où P_{obs} correspond à la proportion d'accord observé (concordance observée) et P_{att} à la proportion d'accord attendu (concordance aléatoire).

Soit $K = 0.50$ pour l'accord entre les juges 1 et 2

Le tableau 2.7 rend compte de l'accord entre tous les experts selon le cœfficient Kappa.

codeur 1+2	codeur 1+3	codeur 1+4	codeur 2+3	codeur 2+4	codeur 3+4	moyenne
P_{obs}=0.91	P_{obs}=0.92	P_{obs}=0.89	P_{obs}=0.90	P_{obs}=0.87	P_{obs}=0.88	K=0.42
P_{att}=0.82	P_{att}=0.85	P_{att}=0.79	P_{att}=0.85	P_{att}=0.79	P_{att}=0.82	/16746
K=**0.50**	K=0.42	K=0.47	K=**0.35**	K=0.39	K=0.37	paires

TAB. 2.7 - *Degré d'accord entre les juges (cœfficient Kappa)*

L'*accord observé* (mesure P_{obs} dans le tableau 2.7) entre chacun de ces juges se situe entre 87 et 92 %. Le cœfficient Kappa indique un *taux d'accord* de jugement entre nos experts situé entre 0.35 et 0.50.

Étant donné la forte disproportion des classes (obsolescence *vs.* non-obsolescence), ces résultats tendent à nous informer plus sur l'accord entre les jugements de non-obsolescence que ceux concernant l'obsolescence : forts P_{obs} et P_{att} liés à D sur-représenté par rapport à A (2397 accords pour la non-obsolescence contre 148 cas d'accords d'obsolescence, *cf.* tableaux 2.3 et 2.4).

Le taux d'accord entre les juges est mieux traduit par le cœfficient *r* de Finn[9] (Hripcsak et Heitjan, 2002). Ce cœfficient permet d'aplanir la disproportion des classes en comparant P_{obs} non pas au P_{att} tel que nous venons de le calculer pour le Kappa mais à une situation aléatoire considérant que chaque annotateur a une chance sur deux de déclarer un segment obsolescent (en situation de hasard). Dans notre situation où il y a deux classes à évaluer, cela revient à considérer un P_{att} égal à 0.5.

Les scores pour le cœfficient *r* de Finn varient de 0.75 pour l'accord le plus bas (les codeurs 2 et 4) à 0.83 pour l'accord le plus haut (codeurs 1 et 3).

Les chiffres du tableau 2.8 montrent que le juge 4 est le plus souvent en désaccord avec les autres juges. Le tableau 2.9 confirme cette tendance. À l'opposé, le juge 1 est souvent assez d'accord avec les autres.

[9]Nous utilisons l'algorithme existant dans le logiciel R (`http://bm2.genes.nig.ac.jp/`

Taux d'accord	codeurs 1+2	codeurs 1+3	codeurs 1+4	codeurs 2+3	codeurs 2+4	codeurs 3+4
Variable *obsol*	0.82	**0.83**	0.78	0.80	**0.75**	0.77

TAB. 2.8 - *Degré d'accord entre les juges deux à deux (cœfficient r de Finn)*

Taux d'accord	codeurs 1+2+3	codeurs 1+2+4	codeurs 2+3+4	codeurs 1+2+3+4
Variable *obsol*	0.82	0.78	0.77	0.79

TAB. 2.9 - *Degré d'accord entre les juges trois par trois et les quatre ensemble (cœfficient r de Finn)*

Ces résultats (Kappa et *r* de Finn) montrent tout d'abord qu'il n'y a pas une grande variation de jugement entre les quatres experts sur la nature obsolescente ou non d'un segment. En d'autres termes, cela nous conforte dans l'idée que l'obsolescence est un phénomène qui fait suffisamment consensus pour être automatisé.

Mais ces résultats montrent également qu'il s'agit d'un phénomène difficile à appréhender et que, dans tous les cas, il serait illusoire de penser qu'on pourra mettre au point un prototype idéal. La suite de ce travail consistera à évaluer l'importance et l'implication des 15 à 20 % de désaccord. L'obsolescence est un phénomène qui devra faire l'objet d'une réévaluation précise et plus consensuelle (*cf.* chapitre 9).

L'obsolescence selon les rubriques thématiques

L'annotation manuelle de l'obsolescence a également mis en évidence la forte variation du nombre de segments selon les rubriques thématiques. Le tableau 2.10 récapitule le nombre et le pourcentage de segments obsolescents par rubrique traditionnellement présente dans une encyclopédie. Il montre à quel point le nombre de segments obsolescents peut varier selon la thématique : de 0 %[10] dans les textes traitant d'*Art et Littératures* à presque 30 % pour la *Géographie*.

Cette section a permis de rendre compte du phénomène de l'obsolescence à travers une évaluation quantitative de l'annotation manuelle du corpus : proportion des segments d'obsolescence dans le corpus, taux d'accord entre les juges sur le phénomène recherché et proportion de segments obsolescents selon le type

RGM2/R_current/library/irr/man/finn.html).

[10]Ce qui ne veut pas dire qu'il n'y aura jamais de mise à jour dans les fiches relevant de ce domaine ; ces chiffres rendent compte de la réalité de notre corpus, non pas de la réalité du type encyclopédique en général.

	nombre total de segments	nombre de segments obsolescents	Pourcentage de segments obsolescents
Géographie	1816	503	**27.7 %**
Économie/ Société/ Droit	1916	290	15.1 %
Sciences et Techniques/ Faune et Flore	1527	297	19.5 %
Histoire	1513	123	8.1 %
Sport	525	26	19 %
Arts et Littératures	332	0	**0 %**
Médecine	1720	123	7.1 %
rubrique inconnue	567	146	**25.7 %**
TOTAL	9916	1508	15.2 %

TAB. 2.10 - *Les segments obsolescents selon les rubriques de l'encyclopédie*

des rubriques encyclopédiques. Nous allons maintenant proposer une description fonctionnelle et textuelle des segments d'obsolescence telle qu'elle ressort de l'annotation manuelle.

2.4.2 Nature et forme textuelle des segments d'obsolescence

Les éléments descriptifs présentés dans cette section font écho, à la fois aux discussions avec les experts (Initiales et Larousse) sur la tâche de mise à jour, et aux observations du corpus annoté. Il s'agit principalement de montrer l'intérêt pour notre tâche d'analyser les textes selon une approche variable et granulaire (*cf.* chapitre 1, p. 15).

L'observation approfondie des segments d'obsolescence tels qu'ils ont été relevés par les juges humains permet de distinguer deux types de segments : nous les évaluons selon la nature des informations qu'ils contiennent d'une part, et selon leur forme textuelle d'autre part.

Concernant la nature des informations à mettre à jour, nous opposons les mises à jour de type « réadaptation » aux mises à jour de type « réactualisation ». Lorsqu'un segment nécessite une *réadaptation*, l'information contenue n'est plus vraie ou ne s'est pas vérifiée. C'est souvent le cas lorsque l'auteur fait des prédictions sur un fait ou sur un événement. Dans le cas d'une *réactualisation*, l'information, dans l'absolu, restera toujours vraie mais, dans un contexte éditorial, il s'avère nécessaire de l'actualiser (par exemple, les chiffres du PIB de l'année en cours).

Concernant la forme textuelle de ces segments, nous avons constaté qu'elle

varie de la taille d'un syntagme nominal au regroupement de plusieurs phrases, jusqu'à atteindre la taille d'un paragraphe ou d'une section[11]. Nous distinguons les notions de *segment d'obsolescence minimal* et de *cadre d'interprétation*.

Nous illustrons ces oppositions sur la base d'exemples de segments obsolescents extraits du corpus ENCYCLO annoté manuellement[12].

Des segments à réactualiser et des segments à réadapter

Dans un *segment à réactualiser*, l'information donnée par le segment textuel reste vraie dans l'absolu mais, pour des raisons éditoriales, les événements et dates associés doivent être plus proches du moment de lecture ou de réédition (*cf.* le schéma 2.5, p. 43).

Dans l'exemple 2.10, le rédacteur informe son lecteur que le nombre de nouveaux cas de maladies du travail est de 160 millions en 2001. C'est une donnée qui, malgré l'inexorabilité temporelle, restera toujours vraie. Cependant, en tant que lecteur, nous préfèrerions sans doute les données concernant l'année de référence du lecteur (éventuellement l'année précédente) car ce qu'une encyclopédie est censée apporter c'est un état des faits de la société le plus réel, le plus actuel possible pour un phénomène donné.

> Les maladies liées au travail posent un problème important. L'organisation mondiale de la santé (OMS) estime, en effet, à **160 millions** le nombre de nouveaux cas dans le monde, **en 2001**. La législation dans ce domaine étant souvent mal connue, un grand nombre de maladies professionnelles ne sont pas déclarées. [...]
>
> Source : ATLAS (fiche Médecine - Les maladies du travail)

Exemple 2.10 - *Un segment d'obsolescence de type réactualisation*

Les segments nécessitant une *réadaptation* contiennent, ou bien une information potentiellement fausse au moment de la lecture ou de la réédition, ou bien une information qui ne s'est pas vérifiée alors qu'elle avait été prédite.

Dans l'exemple 2.11, l'auteur émet des hypothèses sur le déroulement de l'année 2004. Or, si l'on considère le repère temporel actuel, soit l'année 2009, la date de la prédiction date de 5 ans : cette phrase entière est donc devenue obsolète.

L'exemple 2.12 rend compte d'un cas où l'auteur fait des suppositions quant à l'issue d'un conflit. Il y expose deux possibilités et leurs conséquences. Ici, il n'y a pas de date mais les temps verbaux permettent de juger du caractère à la fois non accompli et subjectif de l'auteur (le futur « constituera » et le conditionnel « exprimerait »).

[11] Pour des raisons techniques d'implémentation, c'est l'unité phrase qui a été retenue pour délimiter l'obsolescence. Ceci ne remet pas en cause la variabilité de taille du segment d'obsolescence, qui est réelle mais qui reste un phénomène délicat à traiter automatiquement.

[12] D'une manière générale, les exemples encadrés et dont la source est citée explicitement sont des extraits du corpus ENCYCLO et sont de nature obsolescente (sauf si mention contraire).

L'année 2004 devrait être très largement consacrée à tenter de régler ces questions, en fonction, en particulier, des changements politiques dans certains pays et des résultats des élections européennes de juin 2004. [. . .]

Source : ATLAS

Exemple 2.11 - *Un segment d'obsolescence de type réadaptation - 1*

Un échec provisoire ne constituera après coup qu'un retardement de quelques mois, insignifiant aux yeux de l'histoire, alors qu'un échec définitif exprimerait, au moment de la réunification politique du continent européen, la fin du processus d'intégration imaginé par les « pères fondateurs ». [. . .]

Source : ATLAS

Exemple 2.12 - *Un segment d'obsolescence de type réadaptation - 2*

Les segments d'obsolescence minimaux et les cadres d'interprétation

Un *segment d'obsolescence minimal* est la plus petite unité susceptible d'être modifiée. Il peut s'agir d'un chiffre, d'une date, etc.

L'exemple 2.10 (p. 56) comprend deux segments minimaux : « 160 millions » et « en 2001 ».

Par contre, dans les exemples 2.11 et 2.12, les segments minimaux sont de la taille de la phrase entière : l'information de la phrase entière, perçue dans sa globalité, mérite d'être vérifiée.

Cette notion de segment minimal est directement liée à la quantité d'information à mettre à jour dans le segment en question, qu'il s'agisse d'une réactualisation ou d'une réadaptation. Il semble qu'une réactualisation entraînera le plus souvent la mise à jour de segments courts alors qu'une réadaptation sera plus souvent liée aux segments à gros grains.

Nous avons par ailleurs constaté que, dans bon nombre de cas, il ne suffit pas de connaître uniquement le segment minimal pour être en mesure de procéder à la mise à jour de l'information. Au contraire, il est souvent nécessaire de disposer d'un contexte d'apparition plus large que le segment seul. Nous proposons la notion de *cadre d'interprétation* que nous définissons comme :

- un segment textuel de longueur indéterminée,
- qui présente une certaine homogénéité sémantique,
- qui peut contenir des segments ne nécessitant pas de mise à jour,
- et qui contient au moins un segment d'obsolescence minimal et des indices/-marqueurs de l'obsolescence.

En d'autres termes, il s'agit de fournir un contexte d'interprétation suffisant pour que les mises à jour puissent se faire correctement.

Dans l'exemple 2.10 (p. 56), l'expression chiffrée « 160 millions » est liée au syntagme nominal « le nombre de nouveau cas dans le monde » dans la même

phrase. Ce SN est une reprise anaphorique partielle du SN de la phrase précédente « Les maladies liées au travail ». Il semble dans ce cas nécessaire de disposer *a minima* de ces deux phrases pour être en mesure de procéder à la mise à jour du segment. Un travail sur les chaînes de référence pourrait permettre d'améliorer le repérage automatique des cadres d'interprétation pour l'obsolescence et de faciliter l'interprétation et la mise à jour pour le rédacteur : l'existance d'une chaîne référentielle sur plusieurs phrases rendrait compte de l'existance d'un segment homogène thématiquement au sein duquel des mises à jour doivent être menées. Une autre piste de recherche consiste à considérer les phrases réunies par une chaîne de référence comme un segment à part entière, une unité d'analyse valide au même titre que l'unité phrase. L'ensemble de la méthodologie mise en place dans cette thèse consistant à rechercher des configurations d'indices serait alors menée non pas sur la phrase mais sur ce type de segment textuel regroupant plusieurs phrases réunies par une chaîne référentielle.

Dans d'autres cas, les adverbiaux à l'initiale de la proposition (*cf.* les introducteurs de cadre, p. 98) sont susceptibles d'initier de tels cadres d'interprétation en propageant leur portée sur un ensemble de propositions. Ils donneraient ainsi une unité sémantique cohérente au segment textuel en question. L'exemple 2.13 en est une illustration : les deux propositions suivant celle qui est initiée par l'adverbial spatial « en France » sont sous sa portée sémantique.

§ Les théories keynésiennes [...] **Ainsi, en France,** les chiffres officiels avancés par les autorités sanitaires ne reflètent pas la réalité car ils concernent uniquement les cas indemnisés. Pour autant, la situation s'améliore et on note une augmentation régulière des cas déclarés depuis le début des années 1990. [...]

Source : ATLAS (fiche Médecine - Les maladies professionnelles

Exemple 2.13 - *Un cadre d'interprétation : la portée des IC*

Même si le segment d'interprétation constitue une réalité pour les rédacteurs - utilisateurs potentiels d'un outil de recherche de l'obsolescence, ce travail de thèse ne vise pas sa détection, ni celle du segment minimal d'ailleurs. Nous avons fait le choix de nous baser sur l'unité phrase, principalement pour des raisons techniques (nous y revenons dans la section 6.3, p. 168).

2.5 Conclusion

Dans ce chapitre, nous avons présenté notre corpus [ENCYCLO] comme une sélection de textes issus du monde professionnel qui doivent faire l'objet de mises à jour. Ces textes de type encyclopédique sont très variables quant aux thématiques abordées.

La constitution de ce corpus a permis de se donner les moyens de comprendre la tâche de mise à jour à travers : *(i)* des discussions avec les partenaires (définition globale de la tâche pour permettre de procéder à une annotation manuelle

des textes) et *(ii)* l'annotation manuelle systématique du corpus (environ 10 000 phrases en tout).

L'annotation manuelle du corpus renvoie des informations quantitatives (proportion de segments obsolescents dans l'ensemble du corpus, en fonction des rubriques, accord entre les juges sur la perception de la notion d'obsolescence) et des informations de type descriptif (nature de l'obsolescence, forme textuelle des segments d'obsolescence). Cette annotation manulle permettra également de procéder à l'évaluation du système telle que présentée au chapitre 9.

Les oppositions entre réadaptation / réactualisation et segment minimal / segment d'interprétation permettent de distinguer l'exploitation humaine du système visé (*i.e.* indiquer le type de mise à jour pour le rédacteur pour faciliter son travail de mise à jour) du repérage concret, de l'identification des segments (*i.e.* des portions textuelles variables dans leur taille). Nous verrons dans la troisième partie que prendre ces oppositions en compte est techniquement délicat, du moins dans un travail exploratoire tel que celui-ci : l'obsolescence n'est pas définie en termes de réadaptation / réactualisation par les experts et ce sera probablement un point à développer ultérieurement. Par ailleurs, nous considérons la phrase comme unité d'analyse de l'obsolescence et non le segment minimal ou le cadre d'interprétation, même si cette opposition nous semble intéressante.

Nous avons décrit nos intuitions quant à l'importance du temps, de l'expression de la subjectivité de l'auteur mais aussi d'objets linguistiques discursifs comme les titres (et notamment leur incapacité à accueillir l'obsolescence mais leur capacité à la prédire), les adverbiaux en initiale qui ont un rôle cadratif ou encore l'importance des expressions de mesure. Après un bilan de la première partie, nous proposons une description détaillée des expressions linguistiques susceptibles de permettre le repérage automatique des segments d'obsolescence.

Bilan de la première partie

L'obsolescence n'est pas un phénomène linguistique. Il est créé à travers une situation précise, la mise à jour éditoriale. Pour comprendre et caractériser ce phénomène extra-linguistique, un travail approfondi de description des textes réels et attestés est incontournable et nécessaire.

Notre objectif appliqué, repérer automatiquement les portions textuelles à mettre à jour, s'inscrit dans un contexte applicatif et technique large : les systèmes de gestion et de traitement de l'information et le T.A.L. Nous avons décrit en quoi les applications, les techniques de T.A.L. sont pertinentes pour notre tâche et dans quelle mesure les méthodologies sont ré-exploitables.

Nous avons notamment insisté sur la nécessaire prise en compte de la variabilité du grain d'analyse, de la prise en compte d'indices linguistiques diversifiés pour rendre compte du phénomène complexe qu'est l'obsolescence. Le contexte scientifique et technique dans lequel nous nous inscrivons permet d'envisager des solutions concrètes pour le repérage des segments d'obsolescence.

Qu'est ce que précisément l'obsolescence ? Qu'est ce qu'un segment d'obsolescence ? C'est à travers l'étude d'un corpus de textes issus du monde professionnel que apportons des éléments de réponse à cette question. Les textes du corpus [ENCYCLO] sont annotés par des experts : une évaluation tant qualitative que quantitative de l'obsolescence dans les corpus encyclopédiques (globalement et par rubrique) est menée. Nous validons également la tâche sur la base de mesures sur l'accord entre les experts. Ce corpus annoté manuellement est ainsi le socle pour une description linguistique fine des segments contenant une information susceptible d'évoluer dans le temps.

L'obsolescence (même s'il s'agit d'un phénomène relativement flou et encore mal délimité : l'accord inter-juges n'est pas parfait) apparaît comme un phénomène qui peut être appréhendé par des outils linguistiques. Nous avons proposé une description fonctionnelle en termes de nature des segments (réactualisation et réadaptation) et de forme textuelle (segment minimal et cadre d'interprétation). Ces distinctions montrent qu'un système à granularité variable, adaptable selon les besoins des utilisateurs doit être visé. Les modèles linguistiques et discursifs doivent alors être souples et modulables.

Maintenant que nous avons décrit les segments d'obsolescence et montré que ce sont des objets textuels variables tant dans la nature de la mise à jour que dans leur forme et leur taille, nous allons nous pencher sur les indices linguistiques et

discursifs qui, au terme de notre observation des corpus annotés manuellement sont apparus pertinents pour le repérage de l'obsolescence.

Nous avons déjà évoqué la question du temps et de l'expression du point de vue du rédacteur dans les segments d'obsolescence. Nous avons également évoqué l'idée d'une prise en compte des titres comme des éléments prédicteurs de segments d'obsolescence ou encore les adverbiaux à l'initiale de phrase comme éléments capables de rendre compte des cadres d'interprétation. La partie suivante propose dans un premier temps la description et le classement des expressions linguistiques susceptibles d'être de bons marqueurs de l'obsolescence et dans un second temps la présentation de modèles linguistiques et discursifs au sein desquels les expressions linguistiques locales peuvent être appréhendées.

Deuxième partie

Étude exploratoire pour une description linguistique des segments d'obsolescence

Chapitre 3

Indices observés dans les segments d'obsolescence

Ce chapitre propose une description des expressions linguistiques fréquemment utilisées dans les segments d'obsolescence. Il est l'aboutissement d'une observation manuelle minutieuse des segments d'obsolescence tels qu'ils ont été annotés manuellement par nos experts (*cf.* chapitre 2.3). Ce travail nous a permis d'approfondir notre intuition concernant le rôle des éléments temporels et celui de l'expression du point de vue de l'auteur dans les segments d'obsolescence.

L'observation des annotations manuelles nous a également amenée à considérer l'importance d'autres éléments linguistiques comme le type de propos relaté, les sigles, les mesures et valeurs chiffrées ou encore des expressions appartenant à un champ lexical spécifique (le géopolitique).

Aucun des éléments linguistiques dont nous faisons mention dans ce chapitre ne semble capable d'induire une interprétation concluant sur l'obsolescence ou non d'un segment textuel. S'il est évident qu'il existe des indices plus « forts » que d'autres (comme certains marqueurs temporels), il est nécessaire de garder à l'esprit que c'est en termes de configurations que les indices linguistiques sont pertinents pour décrire l'obsolescence.

Par ailleurs, ce chapitre (ainsi que le chapitre 4 suivant) ne constitue pas un état de l'art sur les différents phénomènes linguistiques observés dans les segments d'obsolescence. Nous ne cherchons pas à nous situer parmi un ensemble de théories concurrentes sur un phénomène linguistique donné et le choix de présentation théorique des questions temporelle, aspectuelle ou modale est essentiellement fondé sur la possibilité d'opérationnalisation de la théorie ou non. Nous sommes consciente que les phénomènes abordés sont en réalité beaucoup plus complexes que la manière avec laquelle nous les traitons mais notre but n'est pas à ce niveau d'analyse : il s'agit ici de délimiter les différents indices linguistiques qui vont servir de base, d'unité d'analyse pour la recherche des configurations d'indices pour le repérage des segments d'obsolescence.

Avant d'entrer dans le détail des expressions linguistiques rencontrées dans les

segments d'obsolescence, nous présentons la question centrale de l'énonciation.

3.1 L'approche énonciative de l'analyse du discours

L'énonciation est définie par Ducrot et Todorov (1972) comme *« un acte au cours duquel des phrases s'actualisent, assumées par un locuteur particulier, dans des circonstances spatiales et temporelles précises »*. Dans cette approche, le sens des unités linguistiques est relié à des facteurs extralinguistiques comme leur référence ou leur prise en charge par l'énonciateur.

Benveniste (1966), à l'origine de cette démarche propose la définition de l'énonciation comme *« la mise en fonctionnement de la langue par un acte individuel d'utilisation »*. Il développe parallèlement une théorie générale des indicateurs linguistiques grâce auxquels le locuteur s'inscrit dans l'énoncé. Ces indicateurs peuvent être :

- des pronoms personnels qui caractérisent le locuteur et son/ses interlocuteurs,
- les indications de temps et de lieu ayant pour référence le locuteur,
- les autres identificateurs du référent (démonstratifs),
- les termes qui impliquent un jugement moral ou personnel,
- les termes modalisants qui concernent le degré de vérité, de certitude, de vraisemblance.

À la suite de Benveniste, Culioli (1999, p.49) écrit : *« énoncer, c'est construire un espace et un temps, orienter, déterminer, établir un réseau de valeurs référentielles, bref un système de repérage par rapport à un énonciateur, à un co-énonciateur, à un temps d'énonciation et à un lieu d'énonciation »*. Il développe l'idée d'*opérations de repérage* comme les moyens linguistiques situant la *saynète* évoquée par la phrase (Enjalbert, 2005, p. 75). Il distingue deux repères, le *repère spatio-temporel* qui permet de situer l'événement, la notion dans l'espace et dans le temps, et le *repère intersubjectif* qui permet de préciser le point de vue de l'énonciateur par rapport à son dire. Enjalbert (2005) souligne que ces opérations sont principalement portées par des tournures syntaxiques et des morphèmes grammaticaux (marqueurs aspecto-temporels, marqueurs modaux, marqueurs de thématisation et de focalisation, etc.).

Guelpa (1997) résume les formes relevant de l'énonciation en les organisant en classes d'éléments :

- les *particules illocutoires ou particules d'énonciation* : elles portent non pas sur l'énoncé mais sur la relation entre le locuteur et son interlocuteur. Leur absence n'enlève rien au contenu intrinsèque du message. Elles servent à exprimer l'impatience, l'indignation, l'irritation, l'étonnement, la surprise, la joie, etc.
- les *modalisateurs de vérité* : ils permettent au locuteur de porter un jugement sur son énoncé. Les modalisateurs de vérité expriment une certitude, une nécessité, la probabilité ou encore la possibilité ; ils ne peuvent être niés.

- les *appréciatifs* permettent au locuteur de porter un jugement sur le caractère normal ou anormal, positif ou négatif de son énoncé. L'appréciatif porte un jugement sur un fait réel, ou du moins dont il présuppose la réalité. Ce n'est pas le cas du modalisateur qui ne fait qu'asserter.
- les *argumentatifs* orientent le discours vers une conclusion positive ou négative
- les *contractifs* sont utilisés pour établir, maintenir ou rompre le contact.
- les *commentatifs* introduisent un commentaire, une parenthèse dans le discours. Ils peuvent être de type exclamatif, être de la forme verbale, prépositionnelle ou nominale.
- les *déictiques* servent à désigner et à dénommer. Le repère est le locuteur au moment de l'énonciation. Les déictiques peuvent être personnels, ostensifs, spatiaux ou temporels.
- les *verbes de modalisation* expriment un jugement sur le degré de vérité de l'énoncé. Ils peuvent être substitués par un modalisateur
- les *faits de position et d'intonation* : tout changement dans la position normale, non accentuée correspond à une intention particulière qui relève de la mise en relief.

Ces formes présentées ci-dessus sont toutes des formes relevant de l'unité phrase. S'il est indéniable que leur repérage passe par le niveau phrastique, il a été montré que leur portée peut s'étendre au-delà, au niveau du texte. Nous développons ce point dans le chapitre 4 à travers la présentation de l'hypothèse de l'encadrement du discours, du modèle de l'architecture textuelle ou encore de la théorie de la prédiction.

Entrons maintenant dans le détail des expressions linguistiques apparaissant dans les segments d'obsolescence et qui s'inscrivent également dans la prise en compte des marques énonciatives.

3.2 Le temps prédominant

Deux types d'indices temporels semblent pertinents pour déterminer l'obsolescence des segments. D'un côté, les expressions déictiques sont à mettre à jour lorsque le moment de lecture est trop éloigné du moment de rédaction : l'interprétation de la *deixis* est dans ce cas perturbée car elle peut soit référer au moment de rédaction soit au moment de lecture. D'un autre côté, il faut examiner les marqueurs de temps qui, par rapport au moment de la rédaction, envisagent une vision future, hypothétique d'un événement dans un temps prédit : ce temps prédit au moment de la rédaction coïncide la plupart du temps avec le moment de lecture[1].

[1]Nous ne traitons ni la *temporalité historique* (*i.e.* fictive) ni la *temporalité discursive* (indication du déroulement de l'argumentation développée par l'auteur) telles que (Vuillaume, 2008) les définit. Notre travail ne prend pas sa source dans le cadre de travaux envisageant la temporalité sous l'angle de la progression narrative (Kamp et Rohrer, 1983) ou de l'enchaînement des événements (Allen, 1984).

Weinrich (1973) distingue les notions de *récit* et de *commentaire*[2]. Dans le récit, le locuteur s'exprime dans une attitude neutre et descriptive ; il s'agit du monde narré ou raconté. Les temps généralement utilisés dans le récit sont l'imparfait, le passé simple, le plus-que-parfait, le conditionnel passé, le conditionnel futur. Dans le commentaire, le locuteur n'est pas neutre, il a un point de vue. Il s'agit du monde commenté. Les temps utilisés sont le présent, les futurs I et II, le passé composé[3].

En observant les temps verbaux de nos encyclopédies, on constate que le mode commentaire est le plus représentatif :

- le présent est majoritaire dans notre corpus : 7278 verbes au présent,
- puis vient le passé composé : 2359 verbes au passé composé,
- le futur est relativement faible : 308 verbes au futur (mais nos programmes ne distinguent pas les futurs I et II),
- l'imparfait est utilisé 345 fois et le passé simple 385 fois.

Le monde de l'encyclopédie est un monde réel, mais passé, sur lequel des auteurs donnent leur point de vue, il est un monde du passé qui potentiellement a des incidences sur le monde présent.

Mais cette opposition entre récit et commentaire ne permet pas de comprendre la place et l'importance des adverbiaux temporels et des éléments de type déictique. Suivant notre intuition sur l'importance du temps dans les segments d'obsolescence (*cf.* section 2.3.1, 43), l'usage des expressions déictiques temporelles sont incontournables.

Un déictique est un mot ou une expression liée à une situation d'énonciation particulière pour laquelle un référent (élément de la réalité auquel renvoie un mot) ne peut être défini qu'en relation avec la situation de communication[4]. Dans le cas de la recherche de l'obsolescence, l'interprétation des déictiques est contrainte et complexifiée par des calculs entre le temps de la rédaction et le temps de lecture du texte.

Pour comprendre ce système temporel complexe où la situation d'énonciation doit être mise en valeur, nous avons choisi de nous baser sur le modèle de Vet (1980) (inspiré des systèmes de Reichenbach (1966), d'Imbs (1968), de Martin (1987)). Il permet notamment de mettre en évidence l'opposition entre valeur déictique et valeur anaphorique des expressions temporelles que ce soit pour les adverbiaux temporels ou pour les temps verbaux.

Ce modèle est composé de deux sous-systèmes temporels *1a/1b* et *2*. Les sous-systèmes *1a* (français écrit) et *1b* (français parlé) sont déictiques car ils établissent

[2]Ces notions sont à mettre en parallèle avec celles d'*histoire* et de *discours* chez Benveniste (1966) : le *discours* appelle l'utilisation du présent, du futur I, du passé composé uniquement aux première et deuxième personnes, du passé simple, de l'imparfait et du futur II ; l'*histoire* utilise les formes de troisième personne de l'imparfait du passé simple et du futur II.

[3]Le passé composé est, selon l'auteur, une marque d'énonciation : il indique l'implication du locuteur dans ce qu'il dit et son engagement.

[4]Moment de rédaction ou moment de lecture ? dans notre cas, l'ambiguïté de l'interprétation est inhérente au type d'objet étudié.

une relation au moment de parole (S) tandis que le sous-système *2* est anaphorique parce que les temps de ce système font référence à un antécédent temporel centré sur un élément dépendant d'un antécédent temporel fixé par le contexte (S'). Le système *2* est commun au français écrit et parlé.

C'est le sous-système *1a* qui nous intéresse particulièrement pour ce travail car il met en avant le rôle déictique des temps verbaux et des adverbes temporels.

3.2.1 Le temps des verbes

Les temps du sous-système *1a*[5] sont les suivants :
- passé-simple (R-S), présent (R,S), futur simple (S-R),
- passé antérieur, passé composé, futur antérieur (E-R + aspect accompli : à R, l'événement est terminé)
- futur proche/périphrastique (aspect prospectif : R-E)

Le passé composé n'exprime pas toujours uniquement le passé. Dans l'exemple « Pierre a rempli la bouteille » tiré de Mascherin (2008), l'auteur montre que le passé composé est compatible avec « maintenant » : il peut donc être situé au moment de la parole, ce qui n'est pas le cas avec l'utilisation d'autres temps du passé. En fait, c'est l'état résultatif du passé composé qui se situe dans le présent, voire même dans le futur : « Dans une heure Pierre a rempli la bouteille » (Mascherin, 2008). Si le procès se situe toujours dans le passé, l'état résultatif du procès est dans un autre moment de l'énonciation. On trouve ici une interaction forte entre sens temporel (localisation) et sens aspectuel (structure interne du procès).

Une étude approfondie des données nous permettra de mesurer le rôle du passé composé dans les segments d'obsolescence (*cf.* les résultats et interprétations des données dans le chapitre 7). S'il semble naturel de penser que les temps du passé ne jouent pas un rôle fondamental dans le phénomène de l'obsolescence[6] et que le présent au contraire semble déterminant, nos observations sont mitigées concernant le futur. Notre première intuition était de considérer le futur comme un bon marqueur de l'obsolescence. En effet, le futur est un temps tourné vers l'avenir qui permet d'évaluer les chances de réalisation d'un procès en termes de probabilité et de possibilité. Il convient cependant de rester prudent sur l'usage du futur car nous avons observé une forte utilisation du futur dans des fiches relatant des faits historiques comme dans l'exemple 3.1. Dans cet exemple, le segment n'est, bien entendu, pas à mettre à jour, il s'agit d'un emploi non-hypothétique du futur.

Ce constat rappelle les hypothèses selon lesquelles il y a deux valeurs pour le futur : une valeur temporelle qui fait référence à une information factuelle mais passée (le cas du futur dans le passé ou futur relatif) et une valeur modale qui reflète les notions de prédiction, de doute, d'incertitude.

[5] E : moment de l'événement ; R : point de référence et S : moment de parole ; S' : élément dépendant d'un antécédent temporel fixé par le contexte ; le tiret (-) marque l'antériorité et la virgule (,) la simultanéité.

[6] Il est toutefois possible de considérer que les temps du passé jouent un rôle dans la détermination de la non-obsolescence.

§ Les révolutions libérales de la fin du XVIIIe s., notamment la Révolution française de 1789, **consolideront** le rôle de la propriété privée.

Source : Corpus GUL (fiche Economie - Le capitalisme

Exemple 3.1 - *Un exemple de futur relatif dans un développement historique (segment non obsolescent)*

> *« Contrairement au présent et au passé, le temps futur n'est pas factuel : les situations dont il est question dans les énoncés construits avec les formes du futur ne sont pas des faits réels, mais seulement des prévisions ou des attentes. Aussi ces énoncés ne sont-ils ni faux ni vrais au moment où ils sont produits. Ceci peut donner lieu à des emplois modaux des formes du futur. »* (Laskowski, cité par Vetters et Skibinska (1998))

Dans le cas de l'exemple 3.1 et plus généralement dans les fiches de type historique, c'est le futur à valeur temporelle qui semble le plus utilisé. Au contraire, dans les segments d'obsolescence, les futurs de type modal sont très présents : les événements que le futur affecte ne sont pas situés dans le prolongement objectif et passé du monde réel, ils appartiennent à un autre monde, celui des intentions ou des prévisions subjectives du locuteur (Lyons, cité par Vetters et Skibinska (1998)).

Se pose alors la question de la distinction entre ces deux types de futur. Dans l'obsolescence, c'est la caractérisation modale du futur qui nous intéresse : comment la repérer ? Notre intuition nous laisse penser que la prise en compte du type de rubrique, par exemple histoire *vs.* géographie, ou de la présence de verbes au conditionnel (*cf.* exemple 3.2) pourraient aider à lever l'ambiguïté. Nous laissons à notre corpus et aux analyses quantitatives la charge de nous éclairer sur ce point (*cf.* chapitre 7, p. 175).

§ Il est fortement probable que ces instruments **joueront**, dans un avenir assez proche, un rôle très important dans la recherche sur le cancer. De plus, ils **permettront** d'élaborer des marqueurs pharmaceutiques. Le domaine de la chirurgie **pourrait**, lui aussi, profiter (dès que les accélérateurs **sauront** préserver les tissus) de l'exactitude offerte par ces différents instruments ainsi que de leur capacité à diriger avec une précision quasi-parfaite un rayon laser.

Source : Corpus ATLAS (fiche Science et Techniques - Les accélérateurs de particules

Exemple 3.2 - *Le futur (modal) dans un segment d'obsolescence*

Le conditionnel semble quant à lui très présent dans les segments d'obsolescence (exemples 3.3 et 3.4). En français, il sert généralement à rendre compte d'un fait soumis à une condition, d'exprimer des demandes polies, d'exprimer une hy-

pothèse, ou encore le futur. Le conditionnel est souvent très lié au futur.

Dans l'obsolescence, c'est dans son emploi hypothétique (exemple 3.3) qu'il semble le plus fréquent. Il semble également très utilisé pour rapporter des faits tout en exprimant un doute à leur sujet (exemple 3.4).

§ De même, selon une équipe de chercheurs suédois, il **devrait** être possible d'ici deux ou trois ans d'effectuer des greffes d'utérus.

Source : Corpus ATLAS (fiche Médecine -La recherche médicale

Exemple 3.3 - *Le conditionnel dans un segment d'obsolescence - 1*

§ Il estime qu'il **faudrait** des ressources humaines et financières 10 à 100 fois supérieures à celles dont on dispose actuellement pour enrayer le déclin de la diversité biologique.

Source : Corpus ATLAS (fiche Faune et Flore - Les espèces animales menacées

Exemple 3.4 - *Le conditionnel dans un segment d'obsolescence - 2*

Dans les trois exemples présentés ici (exemples 3.2, 3.3 et 3.4), les verbes au conditionnel sont également caractérisés par le fait que ce sont des verbes modaux (« pourrait », « devrait » et « faudrait ») : dans Laignelet (2006b), nous considérions les modaux comme indices potentiels de l'obsolescence ; cette étude nous a amené à rejeter cette caractérisation et à privilégier le trait *conditionnel* par rapport au trait *verbe modal*.

3.2.2 Les adverbiaux temporels

Le modèle de Vet (1980) propose également de prendre en compte, dans son modèle, les adverbiaux temporels. Ils sont eux aussi distingués selon la relation qu'ils établissent à S ou S' (*cf.* les explications générales du modèle p. 69). Il oppose :

- les adverbes de temps proprement dits : ils précisent la place d'une situation dans le temps ;
- les adverbes de durée ;
- les adverbes fréquentatifs (« souvent », « rarement », « toujours », « quelquefois ») ;
- les adverbes à analyser avec la notion de présupposition (« déjà », « encore »).

Dans les segments d'obsolescence, ces types d'adverbes peuvent tous, à des degrés divers, être présents. Nous allons voir à travers quelques exemples ceux qui sont prédominants.

Les adverbes de temps proprement dits

De nombreuses expressions entrent dans cette classe d'indices potentiels de l'obsolescence : « prochainement », « actuellement », « à l'avenir », « récemment », « aujourd'hui », « désormais », etc. L'exemple 3.5 illustre ces cas.

§ Dans cet esprit, un appareil **actuellement** développé par un consortium franco-suisse devrait **prochainement** voir le jour.

Source : Corpus ATLAS (fiche Sciences et techniques - Les nanotechnologies)

Exemple 3.5 - *Les expressions déictiques temporelles*

Ces expressions ne s'interprètent que si le moment d'énonciation (*i.e.* le moment de rédaction) est connu par le lecteur. Plus le moment de lecture va s'éloigner du moment de rédaction, plus l'information véhiculée devra faire l'objet d'une vérification ou d'une mise à jour. Ainsi, dans l'exemple 3.5, l'appareil développé par le consortium franco-suisse est peut-être terminé.

Les adverbiaux temporels qui réfèrent à une date proche ou coïncidant avec le moment de lecture sont fréquents dans les segments d'obsolescence. Trois classes d'adverbiaux temporels apparaissent représentatifs de l'obsolescence.

Tout d'abord, une classe très productive concerne les cas où la référence temporelle est postérieure au moment d'énonciation et réfère à une période vague. C'est ce que les exemples 3.6 et 3.7 montrent.

§ Cette situation stratégique lui donnera la possibilité d'exercer une influence non négligeable aux niveaux politique et économique **dans les prochaines décennies.**

Source : Corpus Atlas (fiche Géographie - La Turquie)

Exemple 3.6 - *Les adverbiaux temporels déictiques : postériorité par rapport au moment d'énonciation - 1*

§ La recherche médicale est en plein essor et il y a fort à parier que, **dans les années à venir**, les découvertes scientifiques seront de plus en plus nombreuses et fondamentales.

Source : Corpus Atlas (fiche Médecine - La recherche médicale)

Exemple 3.7 - *Les adverbiaux temporels déictiques : postériorité par rapport au moment d'énonciation - 2*

Deuxièmement, les situations où la référence temporelle de l'adverbial coïncide avec le moment d'énonciation (*i.e.* de rédaction) sont fréquentes dans les segments d'obsolescence : il s'agit souvent de fournir au lecteur un état des lieux pour une situation politique, économique ou encore sociale d'un événement ou d'un

fait ; le lecteur s'attend à ce que la référence temporelle soit la plus proche de celle du moment de lecture.

Le segment textuel de l'exemple 3.8 contient un adverbial référant à une date précise, « en 2007 », date qui s'avère être très postérieure au moment de rédaction et qui est donc une prédiction alors que pour nous, la référence est passée (on est en 2009). Dans ce type de cas, il est nécessaire de faire des calculs précis sur les dates.

§ Actuellement, les transistors mesurent 180 nanomètres et ne devraient plus mesurer que 100 nanomètres **en 2007**.

Source : Corpus Atlas (fiche Sciences et Techniques - Les nanotechnologies)

Exemple 3.8 - *Les adverbiaux temporels ponctuels : postériorité par rapport au moment d'énonciation*

Dans l'exemple 3.9, le lecteur préférera sans doute les chiffres du PIB de l'année la plus récente possible plutôt que ceux de l'année 2002. Mais alors, la comparaison exprimée dans la phrase est-elle toujours pertinente ?

§ Le PIB réel a progressé de 4,3 % en 1997 à 7,4 % **en 2002**.

Source : Corpus Atlas (fiche Géographie - L'Afrique Centrale)

Exemple 3.9 - *Les adverbiaux temporels ponctuels : inadéquation entre le moment de rédaction et le moment de lecture*

Enfin, les cas où un intervalle temporel a été ouvert et n'est potentiellement pas refermé sont fréquent dans les segments d'obsolescence. Les exemples 3.10 et 3.11 illustrent ce cas.

§ **Depuis** ont été mis sur orbite plus de 4 000 satellites, parmi lesquels des vaisseaux occupés par des spationautes.

Source : Corpus GLI (fiche Sciences et techniques - L'espace)

Exemple 3.10 - *Les adverbiaux temporels : intervalle temporel inachevé - 1*

§ Les États-Unis entretiennent des relations étroites avec la Géorgie **depuis 2002**, ce qui s'est s'est traduit par une assistance financière (200 millions de dollars en 2004) et par des accords de coopération militaire.

Source : Corpus GUL (fiche Histoire - Géorgie)

Exemple 3.11 - *Les adverbiaux temporels : intervalle temporel inachevé - 2*

Les adverbes fréquentatifs et les adverbes présuppositionnels

D'autres expressions comme « toujours » ou « encore » (dans « ne sont pas encore faits ») sont également très présentes dans les segments d'obsolescence. Leur particularité est qu'il ne s'interprètent déictiquement que lorsqu'ils sont dans un contexte temporel présent ou futur.

L'exemple 3.12 illustre un cas où « toujours » donne des informations sur la nature obsolescente du segment : il est nécessaire d'aller vérifier si effectivement, il n'y a pas des recherches en cours sur le fait de « remplacer des molécules défaillantes afin de soigner un malade ».

§ Dix-sept ans après ces propos visionnaires, la possibilité de remplacer des molécules défaillantes afin de soigner un malade n'est **toujours** pas d'actualité.

Source : Corpus ATLAS (fiche Sciences et techniques - Les nanotechnologies)

Exemple 3.12 - *Les expressions déictiques temporelles - 1*

Avec ces adverbes, il convient de rester prudent car ils ne sont pas forcément associés à l'obsolescence. Par exemple, dans la phrase, « Petit, il allait toujours à la piscine. » l'adverbe toujours n'indique pas que la phrase soit à mettre à jour. Dans les segments d'obsolescence, ce type d'adverbes semble étroitement lié à la présence d'autres marqueurs de temps : dans l'exemple 3.12, on trouve l'adverbial « Dix-sept ans après ces propos visionnaires, » qui permet de situer la scène et de la localiser précisément dans le temps.

3.2.3 Des syntagmes nominaux temporels et des pronominalisations

Les indications temporelles peuvent enfin être suggérées à l'aide de syntagmes nominaux. Il s'agit généralement de noms adjoints d'adjectifs du type « nouveau », « récent », « dernier ». Dans ces cas là (exemples 3.13 et 3.14), la relation déictique entre le moment de rédaction et l'événement est prépondérante, et, comme nous l'avons déjà expliqué, il est nécessaire de vérifier l'intervalle entre le moment de rédaction et celui de lecture.

§ Cette convention de Washington, du lieu où le texte initial fut adopté le 3 mars 1973, lie quelque 161 États, **le dernier en date** (la Libye) ayant adhéré le 28 janvier 2003.

Source : Corpus Atlas (fiche Géographie - La Turquie)

Exemple 3.13 - *Les syntagmes nominaux temporels : pronominalisation temporelle et valeur déictique*

§ **Le dernier Mondial** s'est tenu en septembre 2003 à Gap, en France.

Source : Corpus Atlas (fiche Sport - Les sports aériens)

Exemple 3.14 - *Les syntagmes nominaux temporels : valeur déictique*

3.3 Valeurs aspectuelles

Nous avons observé que, dans les segments d'obsolescence, le procès est souvent en cours de réalisation alors que, à l'opposé, un procès clairement identifié comme étant achevé est au contraire l'indicateur de la non-obsolescence.

Le traitement de l'aspect constitue une problématique complexe en linguistique. Ce travail n'est cependant pas le lieu d'un développement précis sur la question mais, parce que l'obsolescence est caractérisable par un certain nombre de points relevant de l'aspect, nous choisissons de nous fonder sur la classification proposée par Riegel *et al.* (1994) dans la « Grammaire méthodique du français ». Cet ouvrage représente pour nous un ouvrage de référence de qualité.

L'aspect envisage, d'un point de vue interne, le procès sous l'angle de son déroulement interne, indépendamment de toute considération chronologique. Un processus implique en lui-même une durée plus ou moins longue pour se développer et se réaliser. On peut concevoir ce déroulement interne de façon globale ou l'analyser dans des phases successives (de son début à sa fin). Par exemple, le passé simple dans « il voyagea » présente globalement le procès passé alors que dans « il se mit à voyager », le semi-auxiliaire « se mettre à » saisit le procès passé à son début (Riegel *et al.*, 1994, p. 291).

On distingue en français les oppositions aspectuelles suivantes (Riegel *et al.*, 1994, p. 292) :

- opposition *accompli/inaccompli* (*achevé/inachevé*) : à travers les temps du verbe principalement.
- opposition *perfectif/imperfectif* : à travers le sens du verbe principalement.
- opposition *sécant/non-sécant* (*non-limitatif/limitatif, duratif/ponctuel*) : percevoir le déroulement d'un procès.
- opposition *inchoatif/terminatif* : se situe à l'intérieur des limites du procès.
- opposition *semelfactif/itératif* : le procès est unique ou se répète.
- aspect progressif

L'**aspect accompli** envisage le procès au-delà de son terme, comme étant réalisé, achevé[7]. Il n'est *a priori* pas représentatif de l'obsolescence. En revanche, l'*aspect inaccompli* est potentiellement un bon prédicteur de l'obsolescence : il

[7] Ce sont principalement les formes composées (infinitif passé, subjonctif passé, plus-que-parfait de l'indicatif, passé composé) qui illustrent l'aspect accompli car ils présentent un procès parvenu à son terme final, totalement achevé. Les formes composées expriment également l'antériorité (valeur temporelle et non pas valeur aspectuelle).

saisit le procès en cours de déroulement (le repère T' peut se situer en différentes positions entre les bornes initiales et finales) à l'aide des temps simples (infinitif présent, subjonctif présent, imparfait de l'indicatif) qui saisissent le procès en cours de réalisation.

L'**aspect perfectif** envisage le procès dont l'existence complète et véritable est acquise lorsque le procès est parvenu à son terme. L'*aspect imperfectif*, quant à lui, envisage le procès dans son déroulement, sans visée d'un terme final. La nature perfective ou imperfective d'un procès est transmise principalement par le sémantisme propre des verbes et par leur contexte d'utilisation (« Je lis. » est imperfectif alors que « Je lis un roman. » est perfectif). Cette information ne semble pas pertinente pour la caractérisation de l'obsolescence et dans tous les cas, délicate à mettre en œuvre informatiquement.

Avec l'**aspect sécant**, l'intervalle de référence du procès est envisagé sans limites : il est perçu de l'intérieur et découpé en deux parties : une partie réelle nette et une partie virtuelle, floue. L'utilisation de l'imparfait est typique de l'aspect sécant. Avec l'**aspect non-sécant**, le procès perçu est au contraire saisi globalement, de l'extérieur et enfermé dans des limites (une borne finale lui est assignée). Le passé-simple de l'indicatif est représentatif de ce type de procès.

L'**inchoatif** saisit le procès immédiatement à son début alors que le *terminatif* saisit le procès juste avant sa limite finale. Les périphrases verbales typiques de l'inchoatif sont « se mettre à », « commencer à », les verbes en « -ir » (« rougir », « blanchir »), les verbes en « -iser » dérivés de noms (« scandaliser », « caraméliser ») ou d'adjectifs (« familiariser », « moderniser »). Les périphrases illustrant le **terminatif** sont par exemple : « finir de », « cesser de », « achever de », « terminer de ». Les segments d'obsolescence accueillent plutôt les expressions référant à un procès inchoatif comme le montre l'exemple 3.15.

§ Grâce à leurs poids démographiques et économiques, la Chine, et à un moindre degré, l'Inde **commencent à** avoir une influence sensible sur l'ensemble de la région.

Source : Corpus ATLAS (fiche Sociétée - La mondialisation)

Exemple 3.15 - *Les procès inchoatifs dans les segments d'obsolescence*

La négation d'un terminatif dans l'exemple 3.16 peut également être une bonne indication pour le repérage de l'obsolescence.

§ Le trafic n'est donc pas prêt de cesser.

Source : Corpus ATLAS (fiche Faune et Flore - Les espèces animales menacées)

Exemple 3.16 - *La négation d'un procès terminatif*

L'**opposition semelfactif/itératif** développe l'idée qu'un procès peut être unique (semelfactif) ou se répéter un certain nombre de fois de manière discontinue ou

régulière (itératif). Le caractère itératif se manifeste dans les textes à l'aide d'adverbiaux temporels comme « souvent », « quelquefois », « parfois », « rarement », « toutes les semaines », « une fois par an » et avec les temps verbaux du présent et de l'imparfait. On trouve des exemples (*cf.* exemple 3.17) dans les segments d'obsolescence mais l'itération semble malgré tout un phénomène assez rare dans ces segments.

§ Tous les deux ans ont lieu des Championnats du Monde et d'Europe.

Source : Corpus ATLAS (fiche Sport - Les sports aériens)

Exemple 3.17 - *Les expressions de l'itératif*

Enfin, l'**aspect progressif** rend compte du développement progressif d'une action en continu et par degré. Les exemples 3.18 et 3.19 en sont deux illustrations. L'aspect progressif y est renforcé par l'usage d'un conditionnel.

§ Selon ces dernières, la société de l'information serait **en train de** créer une sorte de « gouvernement mondial », qui rendrait l'État obsolète.

Source : Corpus GUL (fiche Société - L'état)

Exemple 3.18 - *Les expressions de l'aspect progressif - 1*

§ Des recherches **sont également en cours** dans le domaine de la défense.

Source : Corpus ATLAS (fiche Sciences et techniques - Les nanotechnologies)

Exemple 3.19 - *Les expressions de l'aspect progressif - 2*

3.4 La modalité : la position du rédacteur face à ses dires

La modalité est ce qui sert, dans la langue, à exprimer l'attitude du locuteur vis-à-vis de son énoncé. Il peut paraître surprenant de lier obsolescence et modalité. En effet, une encyclopédie a pour but d'informer, de décrire des événements, des concepts de la façon la plus neutre possible, le plus objectivement possible. La rédaction d'une encyclopédie devrait tendre vers le *jugement de réalité* (modalité négative) tel que le définit Riegel *et al.* (1994, p. 579) :

> *« [la modalité] peut être explicite comme dans « il est sans doute parti », où la locution « sans doute » marque le degré de certitude que le locuteur confère à son énoncé, ou incorporé au dictum comme dans « je viendrai demain » où le futur envisage le procès sous l'angle de la probabilité. L'absence totale de modalité correspond alors à un « jugement de réalité ».* »

Nous avons malgré cela constaté un nombre important d'expressions de la subjectivité dans le corpus [ENCYCLO] et plus précisément dans les segments d'obsolescence. Cette manifestation subjective est contenue, mesurée mais bel et bien présente. Nous n'avons trouvé aucune expression comme « je pense » ou « à mon avis ». Cependant l'expression linguistique de la subjectivité est très variée comme nous allons le voir dans cette section.

D'un point de vue linguistique, deux types de modalités sont traditionnellement distinguées Riegel *et al.* (1994, p. 580) :

- les *modalités d'énonciation* renvoient au sujet de l'énonciation en marquant l'attitude énonciative du locuteur envers son allocutaire. Elle se traduisent par l'utilisation des types de phrases énonciatifs : déclaratif/assertif pour exprimer une affirmation, injonctif pour exprimer un ordre, interrogatif pour exprimer un questionnement.
- les *modalités d'énoncé* renvoient au sujet de l'énonciation en marquant son attitude vis-à-vis du contenu de l'énoncé (comme la *fonction énonciative* de Jakobson (1963)). Elles expriment la manière dont l'énonciateur apprécie le contenu de son énoncé : évaluations logiques, appréciations (ordre de l'affectif, de l'évaluatif).

Nous faisons l'hypothèse que certaines modalités, qu'elles soient d'énonciation ou d'énoncé, sont des indices potentiels de l'obsolescence.

3.4.1 Les modalités d'énonciation

L'assertion est de loin le type de phrase le plus courant dans notre corpus. Les types interrogatif et exclamatif, relativement peu nombreux, semblent très souvent associés à l'obsolescence probablement du fait de leur utilisation dans des titres principalement comme le montrent les exemples 3.20 et 3.21. Ce constat est à relier à nos hypothèses quant au modèle de l'architecture textuelle (*cf.* chapitre 4.3, p. 105) et à la notion de prédiction (*cf.* chapitre 4.4, p. 111).

Des technologies nouvelles ?
§...

Source : Corpus ATLAS (fiche Sciences et techniques - Les technologies numériques)

Exemple 3.20 - *Les modalités d'énonciation : l'interrogation*

Attention, danger ! :
§ Dix millions de tentatives de suicides par an, dont 10 % sont fatales.

Source : Corpus ATLAS (fiche Médecine - Les maladies mentales)

Exemple 3.21 - *Les modalités d'énonciation : l'exclamation*

3.4.2 Les modalités d'énoncé

Les expressions linguistiques qui rendent compte de l'attitude du rédacteur sur son énoncé sont assez diverses.

Tout d'abord, il y a les expressions qui marquent le **degré d'engagement du locuteur/rédacteur**. Dans les segments obsolescents, il semble qu'il y ait plus de cas où le rédacteur se distancie complètement des propos qu'il tient en citant ses sources d'information. C'est ce qu'illustrent les exemples 3.22 et 3.23. En linguistique, l'**évidentialité** est, d'une façon générale, l'indication de l'existence et/ou de la nature de la preuve, ou du type de témoignage à l'appui d'une assertion donnée.

§ **Selon les sources**, 75 à 85 % des infections par le VIH ont été contractées au cours d'un rapport sexuel.

Source : Corpus ATLAS (fiche Médecine - Le sida)

Exemple 3.22 - *Les modalités d'énoncé : la distanciation du rédacteur*

Sur le divan :
§ 1,2 million de Français sont suivis par un psy, **selon une étude de la Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)**.

Source : Corpus ATLAS (fiche Médecine - Les maladies mentales)

Exemple 3.23 - *Les modalités d'énoncé : un marqueur évidentiel*

Certaines structures impersonnelles comme « il est urgent de », « il est indispensable de », « il est possible/probable/certain », « il est nécessaire/utile/indispensable » impliquent une **valeur de type déontique** aux propos tenus. Ces expressions expriment la valeur de vérité de l'énoncé en termes d'obligation morale (obligation, interdiction, permission). Les exemples 3.24 et 3.25 illustrent ces cas.

§ **Il est fortement probable** que ces instruments joueront, dans un avenir assez proche, un rôle très important dans la recherche sur le cancer.

Source : Corpus ATLAS (fiche Sciences et Techniques - Accélérateurs de particules)

Exemple 3.24 - *Les modalités d'énoncé : l'usage de déontique - 1*

Dans l'exemple 3.26, la valeur déontique de l'énoncé est fournie à travers l'usage de la locution adverbiale « très certainement ».

Si nous n'avons trouvé aucun exemple de l'usage de verbes exprimant un sentiment, une perception, une opinion, un jugement de vérité à la première personne, nous avons remarqué la présence d'adverbiaux exprimant un **commentaire du locuteur sur son énoncé** comme l'illustre l'exemple 3.27. L'expression « À vrai dire » engage le lecteur à penser qu'il s'agit là du point de vue personnel de l'au-

§ Dès lors que le corps est doté de ces facultés homéostatiques (tendance de l'organisme à stabiliser ses différentes constantes physiologiques), **il est possible** d'affirmer qu'un ostéopathe ne guérit pas un patient mais qu'il aide l'organisme de ce dernier à le faire.

Source : Corpus ATLAS (fiche Médecine - Ostéopathie

Exemple 3.25 - *Les modalités d'énoncé : l'usage de déontique - 2*

§ Les accélérateurs de particules participeront **très certainement** et avec efficacité à la protection de l'environnement dans la mesure où ils seront bientôt capables de transmuter des déchets lourds, c'est-à-dire de leur ôter leur caractère polluant et de les rendre inoffensifs.

Source : Corpus ATLAS (fiche Sciences et Techniques - Accélérateurs de particules)

Exemple 3.26 - *L'expression de l'affectivité et de l'évaluation - 4*

teur : en tant qu'information subjective, elle mérite d'être considérée comme potentiellement obsolescente.

§ **À vrai dire**, dès les années 1960, la coupure entre Flamands et francophones domine la vie politique belge.

Source : Corpus GUL (Entrée Belgique - Histoire)

Exemple 3.27 - *Les modalités d'énoncé : les adverbiaux exprimant un commentaire*

Que ce soit au niveau de la phrase simple, de la phrase complexe ou de l'enchaînement des paragraphes dans le texte, les connecteurs établissent entre les éléments reliés une relation logique et une nuance de sens précise (opposition, cause, conséquence, temps, condition, opposition, comparaison, but, etc.). Ils mettent en place la **structure argumentative des textes**.

Mais les éléments argumentatifs ne sont pas seulement des mots-outils servant l'argumentation logique, ils témoignent en même temps de l'attitude de l'énonciateur face à son énoncé (Sini, 2005). Il peut chercher à aider son lecteur dans un cheminement argumentatif spécifique en utilisant des connecteurs logiques ou au contraire le laisser mettre en place son propre raisonnement sans marque explicite. Nous avons constaté que pour quelques argumentatifs, leur usage était fréquent dans les segments d'obsolescence. Tous ne sont pas pertinents pour l'obsolescence : les plus caractéristiques sont ceux qui introduisent une exemplification, l'illustration d'un phénomène décrit de manière globale. Ainsi dans l'exemple 3.28, l'expression « par exemple » introduit une liste de technologies numériques pour lesquelles il faut vérifier si l'intérêt technologique est toujours valable.

Dans des textes de type encyclopédique, nous ne nous attendions pas à trouver

§ Les métiers de l'informatique éditoriale revêtent un rôle de plus en plus important en raison de la généralisation de la P.A.O. et des méthodes qui lui sont liées (**par exemple**, application des langages SGML, HTML et XML).

Source : Corpus GUL (Entrée Édition - Histoire)

Exemple 3.28 - *Les modalités d'énoncé : l'argumentation*

des marques explicites exprimant le **point de vue subjectif de l'auteur**. Nous avons pourtant relevé de nombreux cas où il est fait usage de noms, adjectifs et adverbiaux affectifs ou évaluatifs dans les segments d'obsolescence.

L'exemple 3.29 montre le cas de l'adverbe « malheureusement » qui exprime le point de vue (négatif) de l'auteur sur un phénomène (en l'occurrence, l'utilisation de la géothermie comme source d'énergie).

§ **Malheureusement**, les sites permettant l'exploitation d'une eau très chaude à une profondeur raisonnable sont extrêmement rares, et la puissance géothermique mondiale installée n'est que de 6 000 MW (dont 45 % aux États-Unis), alors que les possibilités sont estimées à 300 000 MW.

Source : Corpus ATLAS (fiche Sciences et Techniques - Les énergies alternatives)

Exemple 3.29 - *L'expression de l'affectivité et de l'évaluation - 3*

Comme pour l'aspect, le traitement de la modalité est un phénomène extrêmement complexe que nous exploitons pour rendre compte de certains cas d'obsolescence. Notre usage et la classification proposée ici sont superficiels et mériteraient largement d'être développés. Il nous semble cependant incontournable d'évoquer la modalité quand on parle d'obsolescence même si le traitement proposé ne va pas dans le fond des choses. En effet, la modalité permet d'envisager un traitement des énoncés en termes de valeur de vérité, de nécessaire / possible / probable, de jugement de valeur, ce qui est fondamental pour l'obsolescence.

3.5 La question du référent dans les segments d'obsolescence

L'observation du corpus annoté nous amène à considérer des éléments comme les sigles, les expressions de lieu ou encore les expressions numériques. Ils ont comme point commun d'être, dans la plupart des cas, le référent de la phrase, ou de constituer le propos d'un autre référent.

La particularité de ces éléments est que, utilisés seuls ou dans un contexte passé, il ne permettent pas de dire qu'un segment est obsolescent, en revanche, as-

socié à une temporalité ou une modalité future ou hypothétique, ces éléments nous semblent pertinents pour renforcer le jugement d'obsolescence d'une information.

3.5.1 Les sigles, les noms d'organisation et les noms de marque

Un sigle est constitué d'une ou plusieurs initiales servant d'abréviation. Il peut être épelé ou lu (c'est alors un acronyme). Le procédé de siglaison, pratique et économe, est utilisé pour référer à toutes sortes d'entités du monde : organisations, entreprises, noms de personnes, formations politiques, concepts complexes ou innovants, publicité et marketing, etc. Dans la plupart des cas, ce qui motive son utilisation, c'est à la fois le principe d'économie linguistique associé à des raisons socio-culturelles : besoin d'économie d'espace dans les textes écrits, lourdeur/-longueur du nom de certaines organisations/institutions, besoin d'une communication rapide et efficace entre experts ou individus d'une même catégorie socio-professionnelle.

Dans les segments d'obsolescence, il sont souvent associés à une information factuelle précise et locale et qui semble très dépendante d'une situation économique, politique ou sociale particulière.

En français, les sigles sont omniprésents dans les journaux, les publications scientifiques, les émissions télé ou dans la publicité. On constate un nombre important de telles expressions dans les encyclopédies, reflets des usages d'une société.

Souvent associé à un référent unique, le sigle se comporte comme un nom propre[8]. Mais contrairement aux noms propres de personnes, les sigles semblent apparaître fréquemment dans des segments d'obsolescence (exemples 3.30 et 3.31).

§ En France, l'**Institut national de la statistique et des études économiques** (**I.N.S.E.E.**) entérine la définition du **B.I.T.** pour le dénombrement des chômeurs lors des recensements de la population (tous les sept ans) et lors de l'« enquête-emploi » annuelle, qui se déroule habituellement au mois de mars.

Source : Corpus GUL (fiche Économie - Chômage)

Exemple 3.30 - *Les sigles - 1*

§ Il est réparti de la sorte : un peu plus de la moitié revient à l'**INSERM**, environ 10 % au **CNRS** tandis que le reste est distribué aux autres centres.

Source : Corpus ATLAS (fiche Médecine - La recherche médicale)

Exemple 3.31 - *Les sigles - 2*

Les noms de marque ont un fonctionnement similaire (exemple 3.32).

[8]Du moins dans un premier temps car il arrive qu'un sigle devienne un nom de la langue à part entière : le mot « ovni » s'accorde dorénavant en nombre.

§ À Bruxelles, la Direction générale de la concurrence a été appelée à rendre son avis : la Commission européenne a tranché en autorisant **Hachette** à ne conserver que 40 % des actifs de **Vivendi Universal Publishing** (devenu entre-temps **Editis**) qu'il avait acquis.

Source : Corpus GUL (fiche Histoire - L'édition)

Exemple 3.32 - *Les noms de marque*

3.5.2 Les expressions spatiales

Les expressions spatiales ne nous semblent pas représentatives de l'obsolescence. Nous les mentionnons parce que nous avons observé un comportement particulier et intéressant à mettre en valeur. Les adverbiaux spatiaux en position initiale de phrase notamment ont tendance à introduire des exemples, des illustrations venant conclure sur la description d'un phénomène donné. Ils semblent fonctionner comme des adverbiaux antéposés détachés du type « par exemple ». C'est donc plus leur fonctionnement textuel qui entre en jeu que leur fonction idéationnelle : nous développons ce point dans la section 4.2 (p. 98).

§ **À Tchiatoura,** l'exploitation d'un des plus grands gisements mondiaux de manganèse (aujourd'hui en voie d'épuisement) a été à l'origine du développement industriel.

Source : Corpus GUL (fiche Géographie - Géorgie)

Exemple 3.33 - *Les expressions spatiales*

3.5.3 Des mesures, des valeurs chiffrées

L'utilisation de valeurs chiffrées est fréquente dans les segments d'obsolescence. Il s'agit de considérer les expressions référant à des pourcentages ou des valeurs monétaires ou toute valeur numérique présente dans les textes : deux sous-classes émergent, une première qui contient des mesures susceptibles d'évoluer fortement (comme « hab. », « euros », « % » , « ‰ » etc.) et une seconde qui contient celles avec un potentiel évolutif faible (« km2 », « kg », etc.).

Dans les segments d'obsolescence, ce sont les expressions numériques de type évolutif qui apparaissent le plus liées à l'obsolescence.

3.5.4 Les superlatifs introduisant des valeurs chiffrées

Dans la continuité des valeurs chiffrées, nous avons constaté que l'emploi des superlatifs était lui aussi fortement utilisé dans l'obsolescence. Par ailleurs, les expressions superlatives sont très fréquemment associées à une expression de mesure.

§ Le taux de natalité, encore situé entre **20 et 25 ‰** jusqu'aux années 1960, est tombé à **11 ‰** en 2003.

Source : Corpus GUL (fiche Géographie - Géorgie)

Exemple 3.34 - *Les expressions de mesure*

L'exemple 3.35 illustre ce cas.

§ La situation démographique et ses richesses naturelles font du Gabon **le pays le plus riche** de l'Afrique centrale avec un PIB par hab. de 6 200 dollars.

Source : Corpus ATLAS (fiche Géographie - L'Afrique Centrale)

Exemple 3.35 - *Les superlatifs*

3.5.5 Des lexiques spécifiques au domaine

Enfin, nous avons construit une dernière classe qui regroupe des expressions de type « géopolitique ». Cette étiquette sémantique a été créée relativement à notre corpus et au type d'informations spécifiques qui y sont présentes. Il s'agit donc d'une classe purement fonctionnelle et dépendant étroitement de notre corpus d'apprentissage. Nous incluons des expressions comme « densité de population », « taux de mortalité », etc. Ces expressions, comme les superlatifs, ont une forte tendance à annoncer la présence de valeurs chiffrées. Associées à des valeurs temporelles spécifiques (présent ou futur pour l'obsolescence ou passé pour la non-obsolescence), il renforcent alors l'interprétation du segment en entier.

§ Le **taux de chômage** s'est effectivement effondré, mais une partie de la population s'est retirée du marché du travail et de 5 à 10 % des Américains sont marginalisés.

Source : Corpus GUL (fiche Société - L'état)

Exemple 3.36 - *Des lexiques spécifiques au domaine*

Alors que les éléments décrits jusque là sont des éléments linguistiques à part entière, la classe « géopolitique » est dépendante du domaine de référence (le corpus [ENCYCLO]). Il faudra tester l'extension et la portabilité de cette classe sur d'autres textes de type encyclopédique.

Ce sont par ailleurs des expressions qui ont tendance à apparaître dans des positions textuelles saillantes : les titres et les amorces[9] ou encore lorsqu'une valeur

[9]Nous considérons une amorce comme un élément lexical seul situé en début de segment (para-

chiffrée est située dans un contexte proche. Ainsi, il est tout-à-fait possible de constituer automatiquement une telle ressource à partir d'un modèle qui exploiterait ces indices contextuels et typo-dispositionnels (présence dans un titre, en position d'amorce, en position sujet, en cooccurrence avec des valeurs chiffrées, etc.).

3.5.6 Une notion fédératrice : les entités nommées

La plupart des éléments linguistiques décrits dans cette section font écho aux travaux sur la notion d'entité nommée déjà introduite dans la section 1.1.2 (p. 20). Cette notion d'entité nommée ne prend sens que relativement à un besoin applicatif précis et spécifique. Elle permet de regrouper des éléments qui malgré leur diversité linguistique ont un point commun : leur capacité à référer à une entité unique du monde.

Comme le souligne Ehrmann (2008), il est difficile de trouver une définition de la notion d'entité nommée sans qu'elle ne se limite à une *simple* liste des types différents qu'elle recouvre. Sa définition est souvent associée à celle du nom propre avec lequel elle partage de nombreux points. Ainsi, Enjalbert (2005) propose la définition suivante :

> *« [...] toutes les formes linguistiques qui, à l'instar des noms propres, désignent de manière univoque une entité par leur pouvoir de sélectivité : noms de personnes, d'institutions et d'entreprises, de lieux, ainsi que souvent les dates et unités monétaires. »*

Vicente (in Ehrmann (2008)) propose également une définition mettant en avant la relation entre le référent unique d'une entité nommée et celui du nom propre. Il les désignent comme des *« éléments discursifs monoréférentiels qui coïncident en partie avec les noms propres et qui suivent des patrons syntaxiques déterminés »*.

Concernant les expressions linguistiques pouvant être regroupées dans la classe des entités nommées, nous retiendrons trois typologies.

Tout d'abord, celle de G. Bauer (1985) (*in* Daille et Morin (2000)) propose une classification basée sur des considérations d'ordre pragmatique. Il expose cinq catégories principales :

- les *anthroponymes* : personnes individuelles ou groupes (patronymes, prénoms, pseudonymes, gentilés, hypocoristes, ethnonymes, groupes musicaux modernes, ensembles artistiques, orchestres classiques, partis et organisations, clubs sportifs, noms donnés aux animaux familiers) ;
- les *toponymes* : noms de lieux (pays, villes, microtoponymes, hydrotoponymes, oronymes, installations militaires) ;
- les *ergonymes* : objets et produits manufacturés (marques, entreprises, établissements d'enseignement et de recherche, titres de livres, films, publications, œuvres d'art) ;
- les *praxonymes* : faits historiques, maladies, événements culturels ;

graphe, phrase) et suivi de deux-points. Ce point est repris dans le cadre du Modèle d'Architecture Textuelle (*cf.* section 4.3.3, p. 105).

– les *phénonymes* : ouragans, zones de haute et basse pression, astres, commètes.

Cette classification très précise est susceptible d'être valable dans plusieurs domaines d'applications différents. Elle ne prend cependant pas en compte les aspects temporels.

La classification de W. Paik *et al.* (1994) (*in* Daille et Morin (2000)) est constituée de neuf classes principales :

– *classe géographie* : villes, ports, aéroports, îles, comtés ou départements, provinces, pays continents, régions, fleuves, autres noms géographiques ;
– *classe affiliation* : religions, nationalités ;
– *classe organisations* : entreprises, types d'entreprises, institutions, institutions gouvernementales, organisations,
– *classe humain* : personnes, fonctions ;
– *classe documents* ;
– *classe équipements* : logiciels, matériels, machines ;
– *classe scientifique* : maladies, drogues, médicaments ;
– *classe temporalité* : dates et heures ;
– *classe divers* : autres noms d'entités nommées.

Alors que la classification de Bauer met en avant les référents de type humain, cette typologie est orientée vers des référents de type géographique, géopolitique, politique, ce qui nous intéresse particulièrement. Les conférences MUC (p. 21) dont l'objectif est de développer des systèmes d'extraction d'information ont vu le développement de techniques robustes d'identification et de catégorisation des entités nommées. dans le cadre de ces conférences-compétitions, la typologie suivante a été mise en œuvre :

– *classe ENAMEX* : les noms propres qui peuvent faire référence à des noms de personne, de lieu ou d'organisation
– *classe TIMEX* : les expressions temporelles divisées en dates et heures
– *classe NUMEX* : les expressions numériques qui font référence à des pourcentages ou à des valeurs monétaires.

Nous apprécions la simplicité de cette typologie ainsi que sa neutralité vis-à-vis des domaines d'application. Elle permet par ailleurs d'y classer l'ensemble des point développés dans cette section. Ainsi, les classes ENAMEX et NUMEX rassemblent les éléments décrits en 3.5 (p. 81) et la classe TIMEX réunit les éléments des sections 3.2 (p. 67) et 3.3 (p. 75).

Ces classifications montrent à quel point les expressions linguistiques classées sous l'étiquette *entité nommée* peuvent être diverses et variées. Elles mettent également en avant l'importance de l'objectif applicatif pour la caractérisation des entités nommées. Pour autant, la linguistique nous donne les moyens et les outils pour décrire et justifier les éléments des différentes classes proposées.

L'expression « entité nommée » sera utilisée dans ce travail pour faire référence à tous ces éléments linguistiques relativement disparates que nous avons présentées dans cette section 3.5. C'est une notion pratique pour notre tâche car elle nous permet de référer tant aux noms propres de personnes, d'organisation, que des ex-

pressions spatiales ou des unités de mesure.

3.6 Conclusion

Ce chapitre a été l'occasion de présenter les caractéristiques linguistiques et sémantiques fortes observées dans l'étude des segments d'obsolescence annotés manuellement. Le bilan que l'on peut faire de cette présentation est que l'obsolescence est un phénomène sémantique et référentiel diversifié qui le rend relativement complexe à appréhender.

Les aspects temporels sont envisagés à travers le temps des verbes et les adverbiaux temporels. L'aspect est perceptible à travers certaines expressions linguistiques. La modalité peut être soit d'énonciation, soit d'énoncé. Dans ce dernier cas, les notions d'évidentialité, les connecteurs argumentatifs ou encore les adverbiaux évaluatifs ont été décrits. Enfin, les éléments explicitant le référent du discours, soit certains noms propres (sigles, organisations, noms de marque), les adverbiaux spatiaux, les expressions de mesures et les valeurs chiffrées, les superlatifs, ou encore des lexiques spécifiques au domaine, ont finalement été présentés et regroupés sous la dénomination d'entité nommée. C'est dans le chapitre 5 que l'outil de repérage automatique des indices est présenté (outil ALIDIS).

Dans le chapitre 4 suivant, nous présentons les modèles linguistiques qui posent un cadre théorique pour, d'un côté, envisager certaines des expressions que nous venons de voir dans des segments à granularité variable, et de l'autre, introduire de nouveaux éléments discursifs susceptibles d'être des indices de l'obsolescence.

Chapitre 4

Éléments de linguistique discursive : des modèles opératoires pour considérer la variabilité des grains d'analyse

L'objectif de ce chapitre consiste à mettre en place un cadre théorique permettant de comprendre comment des éléments linguistiques divers (ceux qui ont été présentés dans le chapitre précédent) sont à même de fonctionner ensemble dans un objectif commun. En d'autres termes, nous souhaitons nous donner les moyens de comprendre l'organisation à facettes multiples de ces éléments dans le discours : organisation des expressions à l'intérieur des segments, organisation entre les segments d'obsolescence eux-mêmes et enfin organisation entre les segments d'obsolescence et le reste du discours. Il s'agit de dégager un modèle opératoire pertinent pour la description et le repérage automatique de l'obsolescence.

Pour ce faire, nous faisons appel à différentes théories du discours (Laignelet, 2006c, 2004) : encadrement du discours, modèle d'architecture textuelle (MAT) et modèle de la prédiction. Ces approches, notamment le MAT, suggèrent l'existence d'une équivalence fonctionnelle entre des formes discursives et des formes typo-dispositionnelles. Les phénomènes visuels sont alors indissociables des structures syntaxiques et lexicales. Ainsi, la mise en forme des textes ou la position de certains éléments dans la phrase (l'initiale dans l'encadrement du discours) influencent la compréhension et agissent en quelque sorte comme des performatifs textuels.

Comme pour le chapitre précédent, les modèles que nous présentons ci-après ne permettent pas de juger à eux seuls de la nature obsolescente ou non d'un segment textuel. À travers ces modèles, certaines unités discursives (comme le paragraphe, la position, les titres, etc) peuvent alors être considérées comme des indices linguistiques à part entière. Ils sont alors susceptibles de se combiner à d'autres indices plus « traditionnels » (marqueurs de temps par exemple) et ainsi, ensemble, orienter une interprétation obsolescente d'un segment. En d'autres termes, il

s'agit de délimiter les différents indices linguistiques qui vont servir de base, d'unité d'analyse pour la recherche des configurations d'indices pour le repérage des segments d'obsolescence.

Mais avant de présenter les modèles et hypothèses retenus pour ce travail, une introduction rapide à l'étude de l'organisation des discours et un rappel sommaire de quelques notions liées à l'organisation discursive nous semble importants.

4.1 Étudier l'organisation des discours

Le terme proposé par HoDac (2007), EOD pour « *Étude de l'Organisation Discursive* » recouvre l'idée qu'un texte n'est pas simplement un sac de mots, de phrases, de paragraphes, etc. Il est un objet structuré, organisé, hiérarchisé. Il est constitué de parties, de segments de granularité plus ou moins locale, plus ou moins globale qui s'organisent entre eux selon divers modes.

Ce travail se situe dans une approche fonctionnelle de l'étude de l'organisation du discours. L'objectif d'une telle approche est de mettre au jour des relations entre des traits linguistiques et des fonctions discursives. La *Linguistique Systémique Fonctionnelle* développée par Halliday (1985) est un modèle central de cette approche. Le texte/discours est envisagé selon trois points de vue indissociables les uns des autres : il est à la fois une unité de signification (*unit of meaning*), une unité sémantique (*semantic unit*) et une unité fonctionnelle (*« a unit of langage in use »*).

Banks (2005) définit les métafonctions sémantiques à l'œuvre dans les textes, les discours telles qu'elles ont été introduites par Halliday (1985) :

- la *métafonction idéationnelle* est la partie de la signification qui concerne la représentation du monde (participants d'un procès, procès en lui-même, circonstances du procès) ;
- la *métafonction interpersonnelle* concerne les relations établies par le locuteur (relation du locuteur avec ses interlocuteurs et avec son message : théorie de l'évaluation) ;
- la *métafonction textuelle* concerne la signification en termes de l'organisation du texte (structure thématique, structure informationnelle, cohésion).

La composante idéationnelle a été développée dans le chapitre 3 à travers la description des nombreux éléments référentiels qui apparaissent de manière prépondérante dans les segments d'obsolescence (aspects temporels, sigles, mesures, élément spatiaux, etc.). La composante interpersonnelle a également été étudiée à travers notamment les expressions du point de vue de l'auteur.

Dans ce chapitre, c'est la métafonction textuelle qui nous intéresse. Elle met en avant les principes d'organisation du texte. Il s'agit, comme le définit également la théorie de la compositionnalité holiste, de prendre en compte deux principes fondamentaux qui peuvent s'appliquer aux textes (Gosselin et Person *in* Enjalbert (2005, p. 189)) :

- le *principe de compositionnalité* : la signification du tout est déterminé par celle de ses parties,

– le *principe de contextualité* : la signification des parties est déterminée par celle du tout dans lequel elles se trouvent intégrées.

De là découle l'idée d'une *compositionnalité sémantique* permettant la construction du sens par assemblage des divers éléments des structures sémantiques pour obtenir une structure globale cohérente.

Ainsi, nous cherchons à comprendre comment des éléments locaux et *sémantiquement simples* (présentés dans le chapitre 3, p. 65) s'organisent ensemble, dans des structures textuelles diverses, pour permettre l'interprétation d'une information complexe de nature obsolescente.

4.1.1 Texte et Discours

Le terme de *discours* fait l'objet de nombreuses définitions selon les approches, les tendances, les écoles de pensée. Pendant longtemps et même encore dans un emploi familier, il réfère à une production orale. Dans ce travail, ce terme désignera uniquement les énoncés écrits.

D'un point de vue théorique, le *texte* se distingue du *discours*, non dans une relation d'opposition mais en complémentarité : le discours est un texte augmenté de ses conditions de production, de sa situation d'énonciation.

> *« Le texte est un mode d'organisation spécifique qu'il faut étudier comme tel en le rapportant aux conditions dans lesquelles il est produit. Considérer la structuration d'un texte en le rapportant à ses conditions de production, c'est l'envisager comme discours. »* Gravitz, in Sarfati (1997, p. 6)

Ce point est particulièrement important dans notre cas du fait de la nécessaire prise en compte d'informations telles que la date de production des textes, leur situation de production éditoriale, la place que le rédacteur se donne dans ses propres écrits. Le chapitre 3 a montré le rôle de l'énonciation pour la description des segments d'obsolescence.

Un texte est donc un ensemble suivi d'énoncés, il est le résultat d'un acte d'énonciation et constitue ainsi la trace d'un discours ancré dans un contexte. Malgré l'apparente linéarité textuelle, nous envisageons le texte comme un objet complexe et structuré au sein duquel différents mécanismes textuels et discursifs entrent en jeu et rendent compte de relations et hiérarchies entre les différents éléments (ou segments) constitutifs des textes.

Les relations et hiérarchies entre les segments textuels sont en parties expliquées à travers les notions de cohérence et de cohésion.

4.1.2 Cohérence et Cohésion

La cohérence textuelle est définie comme une propriété extrinsèque au texte/discours. Il s'agit d'un concept dépendant étroitement de l'interprétant : le récepteur est amené à construire des relations qui n'existent pas dans l'objet textuel

en question et certaines inférences sont des extrapolations à partir des connaissances du monde que l'on suppose partagées.

> *« Coherence is the quality of meaning unity and purpose perceived in discourse. It is not a property of the linguistic forms in the text and their denotations (though these will contribute to it), but these forms and meanings interpreted by a receiver through knowledge and reasoning. »* (Johnson et Johnson, 1999, p. 55)

Selon Carter-Thomas (2001, p. 35), les considérations sur le contexte et le genre de discours peuvent fortement influencer l'interprétation d'un texte. En d'autres termes, dans une situation discursive précise, au sein d'un genre spécifique, le lecteur aura des attentes particulières qui vont l'influencer dans l'interprétation de la cohérence du texte. De plus, à chaque genre de discours sont associées des stratégies interprétatives qui se manifestent par des instructions intrinsèques. La construction d'une interprétation cohérente passera par la facilité à suivre ces instructions.

S'il est vrai que la façon dont les textes sont perçus est une activité intuitive et étroitement dépendante de chaque interprétant, cela ne signifie pas que cette activité soit un phénomène exclusivement subjectif. Une échelle de la cohérence des textes a été proposée par (Reinhart, 1980) qui classe les textes en trois groupes :

- les textes *explicitement cohérents* : textes qui font le maximum pour faciliter la tâche interprétative du lecteur en respectant les trois principes suivants : la cohésion, la non-contradiction et la pertinence ;
- les textes *implicitement cohérents* : textes qui nécessitent des procédures particulières de la part du lecteur qui va imposer une cohérence dérivée (d'où un travail cognitif particulier) ;
- les textes *incohérents* : textes auxquels on ne peut pas attribuer une interprétation. Il s'agit le plus souvent de textes pathologiques.

La construction d'une interprétation cohérente d'un discours est rendue possible à travers les relations, explicites ou implicites, entre les différents segments d'un discours. Le discours est alors appréhendé à la fois de façon statique et de façon dynamique : statique car les relations concernent les liens et les rapports entre les différents segments du discours ainsi que l'aspect hiérarchique du discours, et dynamique puisqu'elles contribuent à expliquer les processus de construction et de décodage du discours des individus au niveau cognitif.

Les théories qui étudient les relations de cohérence se basent en général sur les marques présentes à la surface des textes[1].

Degand et Sanders (2002) considèrent le texte comme une unité globalement cohérente (*global coherence*) qui présente la particularité de pouvoir être divisée

[1] mais ce n'est pas toujours le cas comme dans la RST (*Rhetorical Structure Theory*) qui se base sur le processus d'interprétation chez les individus et non sur des marques textuelles pour identifier les relations qui interviennent dans le discours.

en segments localement cohérents (*locally coherent*). La cohérence globale est alors fondée sur la relation qu'entretiennent les différents segments localement cohérents.

La structure du discours est ainsi déterminée par ces deux types de cohérence. Afin de délimiter ces segments localement cohérents, Degand et Sanders (2002) affirment que des signaux perceptibles à la surface du texte, les *global discourse markers* (*cf.* section 4.1.3 suivante), fonctionnent comme des instructions au lecteur en l'aidant à construire la représentation mentale du texte la plus adéquate.

Dans *Cohesion in English* Halliday et Hasan (1976) cherchent à faire le lien entre des caractéristiques formelles de la surface textuelle et la qualité globale de cohérence. La cohésion est *« the means whereby elements that are structurally unrelated to one another are linked together, through the dependence of one to the other for its interpretation »*. Les procédés de cohésion permettent de relier une phrase d'un texte à celles qui la précèdent et l'interprétation de cette phrase dépendra en partie des phrases précédentes. La cohésion concerne la manière dont se construit un texte sur la base des relations de signification entre ses éléments.

> *« Cohesion is a semantic notion referring to relations of meaning between elements of a text. »* (Johnson et Johnson, 1999, p. 55)

Halliday et Hasan (1976) expliquent que la cohésion participe de la *texture*. La texture est ce qui distingue un texte de ce qui n'en est pas un. Elle est basée sur les relations de cohésion (*cohesive ties*), c'est-à-dire que l'interprétation d'une unité linguistique (lexicale) dépend directement d'une autre unité se trouvant avant ou après dans le texte. Elle désigne toute l'organisation formelle du texte dans la mesure où cette organisation assure sa continuité sémantique, son isotopie (Sarfati, 1997, p. 28). C'est ce qui donne au texte les propriétés de *« being a text »*.

La cohésion fait référence à la linéarité du texte, à l'enchaînement entre les propositions, aux moyens formels dont dispose le scripteur pour assurer ces enchaînements. Son étude passe souvent par le repérage et l'analyse des marques de relations entre énoncés ou constituants d'énoncés, comme les connecteurs ou les anaphores.

Selon Halliday et Hasan (1976), les différents procédés de cohésion se répartissent en cinq classes d'éléments :

- *la référence* qui peut être endophorique (*via* l'anaphore ou la cataphore) ou exophorique. Elle repose essentiellement sur l'utilisation de pronoms, d'articles démonstratifs et définis, de comparatifs.
- *la substitution* : il s'agit du remplacement d'un élément d'une phrase (nom, verbe ou même proposition) par un autre, en général plus court.
- *l'ellipse* : ici, l'élément de phrase concerné n'est pas remplacé, il a simplement disparu et la cohésion est créée par le caractère volontaire de cette absence.
- *la réitération* : elle comprend la répétition d'une unité lexicale, l'emploi de synonymes, de termes proches (hyponymes et hyperonymes), de termes

généraux ou de collocations (association répétée de deux unités lexicales).

- *la conjonction* qui peut être additive, privative, causale, temporelle ou relevant d'une autre relation logique : elle est réalisée par des expressions comme « mais », « plus tard », « dans ce cas », qui relient deux propositions, deux phrases ou encore deux parties de texte.

Ces procédés de cohésion contribuent à la construction du sens. Il est cependant impossible de s'en tenir à cette seule notion pour expliquer la cohérence. De ce point de vue, la citation suivante n'est pas suffisante :

> *« En somme, la cohésion est l'ensemble des* traces *linguistiques explicites renvoyant à des mises en relation sous-jacentes relevant de la cohérence. »* (Fayol, 1994, p. 111)

Les travaux de Charolles (2002b,a, 2003) entre autres, montrent que les marques de cohésion définies par Halliday et Hasan (1976) ne sont pas les marques exclusives servant de guide aux locuteurs dans l'interprétation des textes et de leur cohérence. Il envisage l'existence de deux grands types de marques de cohésion :

- *la relation de connexion* : des expressions signalent qu'une certaine relation doit être établie entre deux unités adjacentes simples ou complexes. Cette relation peut être référentielle dans le cas de l'anaphore par exemple, ou rhétorique et argumentative dans le cas des connecteurs. Elle donne naissance à des *chaînes*.
- *la relation d'indexation* : des expressions indiquent que plusieurs unités doivent être traitées de la même manière relativement à un critère plus ou moins spécifié par ces expressions. Cette relation est induite par la présence d'adverbiaux cadratifs dont la portée est plus ou moins étendue. Elle donne naissance à des *cadres* (*cf.* section 4.2).

Les marques de relation d'indexation supposent l'existence de segments de discours plus larges que la proposition, dépassant le caractère linéaire des textes et présentant la propriété d'être à la fois structurés localement, mais aussi entre eux afin de participer à la construction d'une interprétation globalement cohérente. Ces marques participent ainsi de la cohérence discursive.

Dans une visée de traitement automatique des textes, envisager que la cohésion est perceptible par des marques explicites est central : aucun programme informatique n'est capable d'interpréter des relations discursives complexes et précises comme peut le faire un être humain. C'est sur la base du repérage d'indices linguistiques particuliers que nous pensons possible le repérage des segments d'obsolescence. Nous gardons cependant à l'esprit que certains segments d'obsolescence ne pourront pas être repérés automatiquement car la référence à l'information évolutive est implicite ou fait appel à des connaissances encyclopédiques qu'un ordinateur n'est pas en mesure d'avoir (par exemple, le fait qu'une guerre se déclare entre deux pays ou qu'un ouragan éclate à l'autre bout de la planète).

4.1.3 Segments, indices et marqueurs

Au début de cette section, un discours a été défini comme un tout unifié, une unité qui tire son sens de l'existence et de la relation entre ses parties. Penser ce tout comme constitué de parties amène à définir la nature et la fonction de ces parties potentiellement constitutives des discours.

Nous emploierons le terme de *segment* pour faire référence à une unité au moins égale à la phrase et susceptible de couvrir un ou plusieurs paragraphes voire une ou plusieurs parties titrées. Un segment est une unité fonctionnelle homogène que ce soit au niveau sémantique, syntaxique, structurel, idéationnel ou/et textuel.

Pour rappel, nous parlons de *segment d'obsolescence* pour faire référence à un segment textuel qui contient des informations susceptibles d'évolution dans le temps.

La notion de *marqueur discursif* est centrale : il s'agit d'une trace linguistique ayant pour fonction d'indiquer les relations qui s'instancient entre des unités composant un discours. De nombreuses théories intègrent la notion de marqueur comme moyen de repérer dans les discours les relations de cohérence. Les *marqueurs d'organisation textuelle* sont des éléments linguistiques pertinents pour la mise en relation des segments. Ce sont des traces textuelles en même temps que des instructions données au lecteur.

Péry-Woodley (2000) définit les marqueurs discursifs comme des *« traces qui constituent une signalisation orientant l'interprétation »*. Elle cite comme exemples de marqueurs discursifs les connecteurs ou *clue words* ou *cue phrases* (« Et », « mais », « pour résumer », « en somme », etc.). Il est possible d'ajouter à cette liste les *marqueurs d'intégration linéaires* ou MIL (« Premièrement », « D'une part [...], d'autre part », etc.), les *introducteurs de cadre* (ou IC[2]) ou encore les titres[3].

Les marqueurs de discours sont des marques linguistiques non référentielles, méta-discursives et dont le rôle est de signaler de façon plus ou moins explicite l'organisation du texte.

Selon Degand et Sanders (2002), afin de signaler au lecteur la cohérence globale du texte, le scripteur a à sa disposition des *global discourse markers* (GDM). Il s'agit d'expressions dont le rôle est pragmatique et dont la fonction est organisationnelle[4]. Les GDM sont de trois types.

Les *metadiscourse markers* réfèrent explicitement à l'organisation du discours.

> *« [Metadiscourse markers are the] explicit references to text boundaries or elements of schematic text structure, either introducing shifts in the discourse or preparing for the next step in the argument »* (Hyland in Degand et Sanders (2002))

Cette classe de marqueurs concerne les expressions utilisées pour le partitionnement (« first », « then », etc.), pour définir des étapes clés du texte (« in sum »,

[2] chapitre 4.2, p. 98

[3] chapitre 4.3, p. 102

[4] *« expressions (...) that are used pragmatically, with a structuring and organizational function »* (Lenk cité par Degand et Sanders (2002))

« to conclude », etc.), pour annoncer des buts discursifs ou de l'énonciateur (« I will argue », « My purpose is », etc.), ou pour annoncer des changements topicaux/thématiques (« so far », « now », etc.).

Les *digression markers* signalent l'introduction d'un topique nouveau et subsidiaire (*push-marker*) ou renvoient au topique principal (*pop-up marker*). Les marqueurs de digression peuvent ouvrir ou clore des séquences digressives de différents types, comme des digressions explicatives (*clarifying digressions*) ou des digressions d'arrière-plan (*background digressions*).

Les *segmentation markers* regroupent des éléments du système linguistique et du système paralinguistique comme la ponctuation, les connecteurs, les adverbiaux, les expressions référentielles qui cooccurrent spécifiquement avec les changement topicaux/thématiques. Ils peuvent être regroupés sous le terme de marqueurs de segmentation. Ils sont moins explicites que les marqueurs métadiscursifs et les marqueurs de digression car ils n'indiquent pas systématiquement comment les segments de discours peuvent être reliés.

Les définitions du marqueur discursif telles qu'elles sont présentées jusqu'à présent supposent une relation entre la forme d'une expression et sa fonction dans le discours. Dans ce travail, il n'existe pas de marqueur spécifiquement lié, dédié à l'obsolescence : il n'y a pas de relation entre une forme précise et une fonction (l'obsolescence).

De plus, un marqueur discursif contribue principalement de la composante textuelle des discours car il indique au lecteur comment relier les éléments d'un discours (rôle instructionnel au lecteur). Dans certains cas, il peut également relever de la composante idéationnelle et de la composante textuelle en cela qu'il délimite des segments textuels spécifiques relativement à une information sémantique particulière : c'est par exemple le cas des introducteurs de cadre ou des titres (*cf.* sections 4.2 et 4.3).

Dans cette idée, nous apprécions les travaux de HoDac (2007) qui distingue les types d'indices suivants :

- les *indices textuels* : ils peuvent être linguistiques et on y range les marqueurs méta-discursifs, les anaphores, etc. ; ils peuvent être typographiques et dans ce cas on traite par exemple les surlignements ou encore les mises en gras ; enfin ils peuvent être propres à la structure du texte, comme les titres.
- les *indices texto-idéationnels* : ce sont par exemple les syntagmes prépositionnels ; ces indices apportent un sens instructionnel et un sens propositionnel (*cf.* l'hypothèse de l'encadrement du discours, section 4.2, p. 98).
- les *indices texto-interpersonnels* : il s'agit par exemple des adverbes modalisateurs, des constructions impersonnelles.

Un marqueur est alors soit un indice isolé contraint, soit une configuration d'indices. C'est la combinaison même de certain indices qui permettra de marquer l'obsolescence.

4.1.4 Envisager les indices en complémentarité

Les indices sont des marqueurs potentiels de l'organisation discursive car ils permettent, à des degrés divers et à des niveaux variés, de rendre compte à la fois de la délimitation de segments mais aussi de la relation et du type de relation que ces segments entretiennent entre eux.

Les travaux sur ces notions (Halliday et Hasan, 1976; Degand et Sanders, 2002; Charolles, 2003; Péry-Woodley, 2000; HoDac, 2007) mettent en exergue l'aspect multiple, multidimensionnel et complexe de l'objet *discours*. Nous insistons notamment sur la problématique centrale concernant la variabilité du grain d'analyse dont nous avons déjà parlé au chapitre 1.3.2 (p. 31). Le chapitre 6 (p. 161) reprend cette question théorique sur le statut de l'unité d'analyse et rend compte des incidences qui se répercutent lorsqu'on cherche à implémenter des traitements automatiques sur le discours.

Dans les segments d'obsolescence, les expressions linguistiques sont de formes très variées : un mot, une expression, un ensemble d'expressions apparaissant dans des phrases successives, les marques typographiques ou encore les titres ou les divisions paragraphiques. Il est fréquent que ces éléments soient traités de façon autonome en établissant une relation binaire entre un élément (ou plus largement une classe d'éléments) et une fonction discursive.

Nous souhaitons pour notre part aller plus loin et nous proposons de penser ces éléments non pas de manière isolée mais en configurations.

> *« C'est d'ailleurs souvent la combinaison de plusieurs marques, réparties un peu partout dans la phrase, qui produit la propriété globale en question. Ainsi, c'est la combinaison de l'imparfait, d'un verbe exprimant un événement ponctuel et un complément de type prospectif qui fait [que la phrase suivante] exprime le mode irréel. »* J. Rebeyrolle (Enjalbert, 2005)

Cette orientation de recherche est relativement récente et les travaux de HoDac (2007); Bouffier (2008); Widlöcher (2008) nous ont permis d'évoluer dans ce sens (*cf.* section 1.3.3, p. 32, section 1.3.2, p. 31). Widlöcher (2008) parle de *faisceaux d'indices*, HoDac (2007) de *configurations d'indices* et Bouffier (2008) de *combinaisons d'indices*.

> *« Notre objectif est en effet de découvrir des procédés de signalement, et non d'examiner le fonctionnement d'éléments définis d'emblée comme des marqueurs. Cette approche nous amène à observer que l'identification d'une structure discursive repose rarement sur un élément lexical isolé mais plutôt sur l'influence conjointe de facteurs multiples de nature parfois autre que lexicale (tels le type de texte, la structure de document, la position de la portion de texte en cours de lecture dans la hiérarchie du document, etc.). »* (HoDac et Péry-Wodley, 2008)

Ainsi, les indices exploités dans le cadre de cette recherche sont à la fois des indices appartenant à des niveaux textuels multiples (niveaux morpho-syntaxique, syntaxique, typo-dispositionnel et discursif) et en même temps, c'est leur combinaison qui va être pertinente et que nous proposons de mettre au jour.

Les sections suivantes décrivent les modèles et hypothèses linguistiques permettant de décrire les segments pertinents pour nos travaux.

4.2 L'hypothèse de l'encadrement du discours

L'*hypothèse de l'encadrement du discours* définit un cadre de discours comme un regroupement de plusieurs propositions sous un critère sémantique véhiculé par une *encadrement du discours, !expression introductrice de cadre* (IC). Un IC est un adverbial situé à l'initiale d'une proposition, généralement en position détachée. Les IC ont pour fonction de signaler que plusieurs propositions apparaissant dans le fil d'un texte entretiennent un même rapport avec un certain critère et sont de ce fait regroupables à l'intérieur d'unités désignées par le terme de *cadre* (Charolles, 1997).

Les cadres de discours sont des *« unités de segmentation originales venant s'ajouter aux unités typo-dispositionnelles (paragraphe, tiret, puces,...) qui sont des sortes de cadres sous-spécifiés sémantiquement (le critère de regroupement des propositions n'étant pas signalé sauf quand il y a titraison) et ils contribuent de ce fait à l'organisation et donc à la cohésion du discours. »* (Charolles et Péry-Woodley, 2005).

La raison pour laquelle nous traitons précisément ces éléments de discours est qu'ils participent des métafonctions sémantiques définies par Halliday et Hasan (1976) (*cf.* p. 90). La position de ces éléments à l'initiale de la proposition ainsi que le fait qu'ils soient peu intégrés leur confère ainsi le double rôle suivant :

- une *fonction textuelle* : des segments discursifs (les cadres) sont mis en évidence par la présence d'introducteurs de cadres qui ont pour fonction de regrouper des segments tels que les propositions ou les paragraphes ou tout autre type de segment ; Charolles et Vigier (2005) parlent de *rôle cadratif* (fonctionnement d'indexation) ;
- une *fonction idéationnelle* (ou représentationnelle) : les introducteurs de cadre posent un critère sémantique suivant lequel les propositions suivantes sont à interpréter ; Charolles et Vigier (2005) parlent de *portée sémantique*.

  > *« [...] les adverbiaux les moins intégrés dans la phrase qui les accueille sont plus ou moins prédestinés quand ils sont antéposés à jouer un rôle dans la structuration du discours et donc à assumer en plus de leur fonction idéationnelle (ou représentationnelle), une fonction textuelle. Cette propension est manifeste avec les adverbiaux connecteurs. Elle est beaucoup moins évidente avec les adverbiaux non grammaticalisés comme les syntagmes prépositionnels spatiaux et temporels qui sont très souvent détachés en tête de phrase et qui*

> *contribuent au contenu idéationnel des énoncés. »* (Charolles et Péry-Woodley, 2005).

Sur le plan syntaxique, une expression introductrice de cadre est un syntagme prépositionnel en position préverbale : cela correspond aux compléments essentiels, et non argumentaux. Ces constituants peuvent être de diverses natures, adverbes, syntagmes prépositionnels ou encore syntagmes nominaux ; ils sont non intégrés syntaxiquement, phrastiques ou non et en position préverbale.

Nous présentons cinq types de cadres : les cadres organisationnels, les cadres locatifs spatiaux et temporels, les cadres thématiques, les cadres médiatifs et enfin les cadres qualitatifs. Nous développons ces notions en les illustrant d'exemples de segments d'obsolescence issus de notre corpus.

Les cadres organisationnels présentent la caractéristique forte de créer une véritable structure textuelle (Jackiewicz, 2005; Jackiewicz et Minel, 2003). Les IC organisationnels peuvent être des *marqueurs d'intégration linéaire* (MIL). Les MIL sont indépendants des contenus sémantiques des segments qu'ils introduisent et qu'ils relient entre eux sur le mode d'une série.

Les IC organisationnels peuvent être également de simples adverbiaux argumentatifs comme l'illustre l'exemple 4.1.

Par exemple, l'hémato-oncologie pédiatrique, une spécialité récente qui traite les enfants atteints de leucémie, a fait ses preuves puisqu'aujourd'hui entre 70 % et 80 % des leucémies aiguës lymphoblastiques (des cellules) guérissent contre 30 % au début des années 1970. [...]

Source : Corpus ATLAS

Exemple 4.1 - *Un segment d'obsolescence initié par un introducteur de cadre organisationnel*

Tous ne sont pas pertinents dans le repérage des segments d'obsolescence : ainsi des MIL tels que « Premièrement [...]. Deuxièmement [...] » semblent ne jamais apparaître au sein d'un segment d'obsolescence dans notre corpus de travail. Certains marqueurs organisationnels seulement sont intéressants comme « par exemple » qui est illustré dans l'exemple 4.1 et que nous avons déjà présenté dans l'exemple 3.28 (p. 81).

Les circonstants de temps ou de lieu (Charolles *et al.*, 2005; LeDraoulec et Péry-Woodley, 2005; Sarda, 2005) détachés en position frontale signalent que les contenus propositionnels rapportés doivent être relativisés à certaines périodes ou certaines zones de l'espace. En plus de leur rôle textuel (regroupement de propositions), ils indexent des contenus exprimés par des phrases qui sont à propos d'autre chose, le plus souvent à propos des référents dénotés par leur sujet (rôle idéationnel) (Charolles et Péry-Woodley, 2005).

Dans l'exemple 4.2, l'introducteur de cadre temporel « Aujourd'hui » étend sa portée sémantique sur l'ensemble de l'extrait. Les valeurs chiffrées données sont ainsi à interpréter relativement à cette période temporelle donnée : ainsi, tout comme la valeur temporelle associée à « Aujourd'hui » évolue en même temps

qu'on avance dans le temps, ces chiffres sont susceptibles d'évoluer également et de devoir subir des mises à jour.

> Aujourd'hui, plus de 25 % de ces transactions s'effectuent ainsi en euros (contre 48 % pour le dollar). Ce chiffre devrait encore progresser dans la mesure où la zone euro représente à elle seule 30 % du commerce international, soit deux fois plus que les États-Unis et que son PIB atteint 80 % du PIB américain.
> [...]
>
> Source : Corpus ATLAS

Exemple 4.2 - *Un segment d'obsolescence initié par un introducteur de cadre temporel*

L'exemple 4.3 montre qu'un introducteur de cadre spatial peut également être utile pour le repérage de segments d'obsolescence. Dans ce cas, l'expression « En France » fournit au lecteur un repère spatial, une sorte de *arrière-plan sémantique* qui permet d'interpréter l'ensemble des informations relativement à ce repère. Ici, l'expression spatiale en elle-même n'a pas à subir de mise à jour : c'est très clairement le rôle cadratif de cet élément qui est utile pour le repérage du segment d'interprétation. La chaîne de référence (identifiable à travers les sujets suivants : « le budget gouvernemental », « Il », « Ce budget ») coïncide parfaitement avec le cadre spatial (Ho-Dac et Laignelet, 2005).

> En France, le budget gouvernemental alloué à la recherche médicale tourne autour de 900 millions d'euros. Il est réparti de la sorte : un peu plus de la moitié revient à l'INSERM, environ 10 % au CNRS tandis que le reste est distribué aux autres centres. Ce budget étant insuffisant pour financer les activités du très grand nombre d'associations [...].
> [...]
>
> Source : Corpus ATLAS

Exemple 4.3 - *Un segment d'obsolescence initié par un introducteur de cadre spatial*

Les cadres thématiques sont généralement utilisés pour signaler au lecteur comment l'auteur organise les informations ou pour mettre en valeur ce qui est important pour lui (Porhiel, 2005). Ils circonscrivent le domaine de validité de la proposition qu'il préfixe au champ d'activité et de connaissances fourni par l'IC (Vigier, 2003).

Les expressions linguistiques qui instancient des cadres thématiques sont du type « au sujet de », « à propos de », « en ce qui concerne », « au chapitre (de) », « concernant », « sur », « quant à », etc. L'exemple 4.4 montre un cas de cadre initié par un introducteure de cadre thématique.

Comme tout IC, les introducteurs thématiques sont syntaxiquement peu soudés, qu'ils préfixent une ou plusieurs propositions et qu'ils partitionnent l'information. Ces expressions sont intéressantes dans la production d'un résumé car ce sont des marqueurs linguistiques indépendants du contenu sémantique des textes qui peuvent donc être réutilisés pour tout type de discours (Ferret *et al.*, 2001).

En ce qui concerne la finance, et malgré l'adoption de l'euro, le dollar reste largement la monnaie internationale dominante ; l'influence politique des grands établissements financiers américains est prépondérante, notamment pour la régulation des services et des marchés financiers. Parmi les pays émergents, les États d'Amérique latine figurent en bonne place, mais leur situation reste fragile (comme le montre la crise argentine) ou très dépendante de l'économie américaine. Le Mexique en particulier, du fait des nombreuses zones franches situées sur la frontière américano-mexicaine, fabrique des produits destinés principalement au marché nord-américain. [. . .]

Source : Corpus ATLAS

Exemple 4.4 - *Un segment d'obsolescence initié par un introducteur de cadre thématique*

Dans **le cas des cadres médiatifs**, l'auteur impose au lecteur une distanciation par rapport aux propos qu'il relate. Ils signalent la façon dont le locuteur a acquis l'information communiquée dans la proposition qu'ils indexent.

D'après les estimations de la Banque des règlements internationaux, le montant des transactions financières est cinquante fois plus important que la valeur du commerce international portant sur les marchandises et les services. L'écart devient gigantesque lorsqu'on rapporte ce montant aux PNB (à titre d'exemple, le PNB de la France - de l'ordre de 140 milliards de dollar en 2001 – équivaut au montant des transactions quotidiennes sur le marché des changes). [. . .]

Source : Corpus ATLAS

Exemple 4.5 - *Un segment d'obsolescence initié par un introducteur de cadre énonciatif*

Dans notre corpus, il s'agit le plus souvent d'expressions référant à un organisme, comme c'est le cas ici ou à un support d'information (par exemple avec « Selon le rapport »). Ce qui est intéressant pour notre objectif applicatif de création d'outil d'aide à la mise à jour, c'est que grâce aux cadres énonciatifs, le rédacteur chargé de mettre à jour l'information reçoit un cadre d'interprétation suffisant lui permettant de savoir dans quelle source vérifier, et si besoin, de modifier l'information.

Les cadres qualitatifs, dans la mesure où ils indexent un point de vue plus personnel de la part du scripteur, co-habitent souvent avec la présence d'un segment d'obsolescence et nécessitent fréquemment une mise à jour de l'information contenue dans le segment.

4.2.1 Conclusion et Positionnement

L'encadrement du discours a été utilisé dans de nombreuses applications de T.A.L. notamment comme contribution aux techniques de segmentation automatique dans le cadre de systèmes de résumé automatique. Ainsi les cadres thématiques ont été exploités par Ferret *et al.* (2001), les cadres organisationnels par

Malheureusement, les sites permettant l'exploitation d'une eau très chaude à une profondeur raisonnable sont extrèmement rares, et la puissance géothermique mondiale insatllée n'est que de 6000 MW (dont 45 % aux Etats-Unis), alors que les possibilités sont estimées à 300 000 MW. Des expérimentations ont lieu, consistant à fissurer, par des explosions souterraines, les roches chaudes entre deux sondages judicieusement choisis, puis à injecter de l'eau par l'un des sondages pour en exploiter la vapeur de l'autre côté. Une station expérimentale fonctionne sur ce principe à Los Alamos, aux États-Unis, mais les résultats restent médiocres.
[...]

Source : Corpus ATLAS

Exemple 4.6 - *Un segment d'obsolescence initié par un introducteur de cadre qualitatif*

Jackiewicz (2002), les univers de discours par Bilhaut (2007).

Le corpus [ENCYCLO] a fourni des illustrations pour la présentation des cadres de discours et des introducteurs de cadres : cela montre l'intérêt de leur prise en compte pour le repérage des segments d'obsolescence et plus spécifiquement des cadres d'interprétation. Ainsi, c'est à la fois leur sémantisme (relativement à l'information qu'ils véhiculent : temporel, spatial, notionnel, etc.) et leur capacité cadrative qui vont nous intéresser.

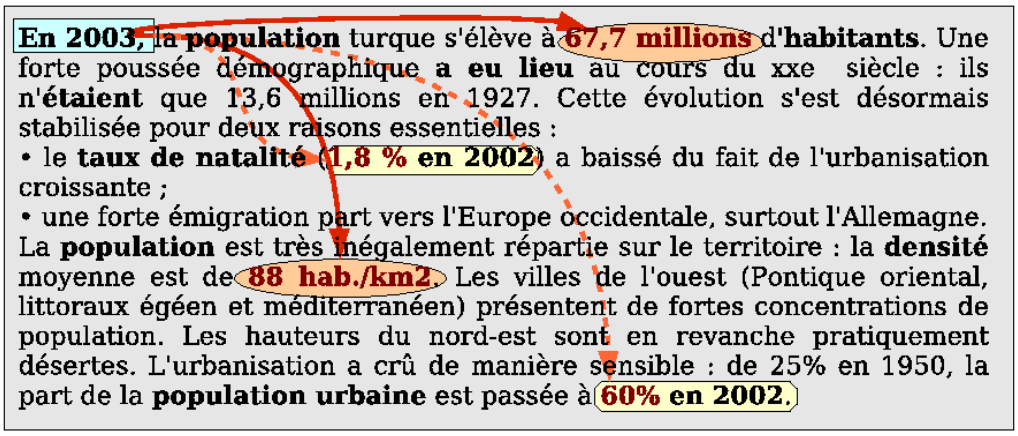
Dans l'exemple 4.7, les éléments à mettre à jour sont relativement locaux puisqu'il s'agit principalement de procéder à la réactualisation des dates et des valeurs chiffrées associées. En effet, la forme générale du paragraphe ne nécessite pas de modification. Ce sont les associations entre la référence temporelle « en 2002 », la valeur chiffrée « 1,8 % » et l'expression « taux de natalité » ou encore entre « En 2003 » et « 67.7 millions d'habitants » ou encore « 68 hab./km2 » qui doivent être mises à jour. Or cette dernière association ne peut être considérée que si la portée sémantique de l'adverbial temporel est reconnue. La capacité textuelle et idéationnelle de l'introducteur de cadre délimite un segment qui recouvre la notion de *cadre d'interprétation* présentée dans le chapitre 3 (p. 57).

Nous allons maintenant nous intéresser à des segments discursifs *à gros grain* : les titres et les paragraphes. Pour cela, nous commençons par présenter le Modèle de l'Architecture Textuelle qui offre un cadre pertinent pour l'étude de tels objets.

4.3 Le Modèle de l'Architecture Textuelle (MAT)

Le MAT[5] est un modèle qui s'applique aux textes écrits. Il considère les critères visuels des textes et permet de représenter l'architecture de ces textes et d'en montrer la cohérence. Le texte y est abordé sous son angle physique, concret : il est

[5]Dans cette section, nous nous sommes inspirée des documents suivants : Luc et Virbel (2001), (Luc, 2000) et Virbel (1985).



Exemple 4.7 - *Les introducteurs de cadres pour le repérage des segments d'interprétation*

un *« énoncé inscrit sur un support matériel »* (Luc, 2000). Nous avons souligné l'importance de la typo-disposition dans les fiches des Éditions Atlas notamment. Le MAT est un modèle opératoire pertinent pour rendre compte de l'importance de certains objets textuels : nous exploiterons principalement les objets titres mais également les aspects positionnels liés à la structure des documents.

Dans le cadre du MAT, la mise en forme matérielle est considérée comme un acte textuel exprimant une intention communicative spécifique qui est mise en valeur.

4.3.1 La notion de métalangage

Le MAT s'inspire de la *théorie transformationnelle* de Harris : Harris pose que la langue peut entièrement être décrite par elle-même et qu'il n'est donc pas nécessaire d'avoir recours à un autre système notationnel que la langue elle-même. Un *métalangage* est donc un langage composé de phrases dont le rôle est métalinguistique. Il permet de décrire la langue (Luc, 2000, p. 35).

Par exemple, les phrases métalinguistiques suivantes (ou *métaphrases*) décrivent la structure grammaticale de la phrase « Max mange un steak » :

(4.1) (a) « Max » est le sujet de « mange » dans la phrase « Max mange un steak »
(b) la phrase « Max mange un steak » comporte 4 mots

Harris démontre que les phrases de la langue possèdent la propriété d'être décomposables et calculables en phrases primaires et que, inversement, il existe un procédé récursif permettant d'engendrer toutes les phrases à partir d'un sous-ensemble fini d'assertions et au moyen d'un ensemble fini d'opérateurs.

Les phrases élémentaires sont des éléments primitifs à partir desquels l'ensemble des phrases de la langue vont être construites.

Les opérateurs de base pour l'anglais sont les suivants :

- les *opérateurs d'expansion de mot* permettent d'introduire des ajouts de base ;
- les *opérateurs verbaux* permettent de dériver un prédicat verbal en un autre ;
- les *opérateurs de phrase* permettent de transformer une phrase en une phrase dérivée ;
- les *opérateurs de connexion* s'appliquent sur un couple de phrase : coordination, subordination, comparatifs, relatifs ;
- les *opérateurs de permutation* permettent la permutation de symboles à certains endroits bien précis de la chaîne phrastique ;
- les *opérateurs de réduction* rendent possibles des effacements importants et complexes sur toute la longueur de la phrase ;
- les *opérateurs de changement morphophonématique* permettent de changer la forme morphophonématique d'un morphème.

Un texte peut ainsi être décomposé en un ensemble de phrases liées entre elles par un ensemble d'opérateurs de base.

Dans le modèle de l'architecture textuelle, le *métalangage architectural*, sous-langage spécialisé, permet de formaliser un certain aspect des structures présentationnelles des textes (Luc, 2000).

4.3.2 Définitions et propriétés des concepts du MAT

Le modèle de l'architecture textuelle étudie la *mise en forme matérielle* des textes comme étant *« l'ensemble des propriétés de réalisation appliquées à un texte »* (Pascual et Péry-Woodley, 1995). Ces propriétés peuvent être :

- *lexico-syntaxiques* : nominalisations, numéralisations, formes interrogatives, etc.
- *typographiques (ou morphologiques)* : habillage du caractère (police, corps, style, couleur, etc). On peut distinguer à ce niveau deux règles de composition : la *microtypographie* (assemblage à l'intérieur d'un bloc) et la *macrotypographie* (assemblage des blocs entre eux).
- *dispositionnelles* : spatialisation sur le support, agencement des objets textuels et non-textuels (justification, colonnage, marges, interlignage, etc).

La **mise en forme matérielle** (MFM) est donc ce qui permet de percevoir la structure d'un texte.

Un **Objet Textuel** (OT) est *« un segment caractéristique de texte, rendu perceptible par un jeu de contrastes de la mise en forme »* (Luc, 2000) (mise en relief, mise en parallèle, etc.). Parmi les exemples courants d'OT, on peut trouver les définitions, les parties ou les titres.

Une **Unité Textuelle** (UT) est *« un segment de texte ne comportant aucun OT, c'est-à-dire un segment de texte entièrement discursif »* (Luc, 2000).

L'**architecture textuelle** est la structure d'un texte, rendue visuellement accessible par sa MFM. Elle est définie comme *« la composante abstraite du texte, constituée de l'ensemble des objets textuels ainsi que des relations qu'ils entretiennent entre eux »* (Pascual et Péry-Woodley, 1995).

4.3.3 La typo-disposition comme indice de l'obsolescence

Dans les documents encyclopédiques, la typo-disposition est très souvent soignée. Elle permet de mettre en valeur des éléments considérés par l'auteur comme importants ou significatifs par rapport au sujet traité.

Le sous-corpus [ATLAS] présente une mise en page très riche[6]. Les éléments importants mis en valeur par ce moyen visuel doivent donc être impérativement mis à jour, d'abord parce qu'ils sont visibles au premier coup d'œil puis parce qu'il s'agit justement souvent d'informations potentiellement obsolescentes. Il s'agit notamment des encadrés à droite dans les fiches Atlas.

L'importance des objets titres dans les segments d'obsolescence

Les titres occupent indéniablement une place caractéristique au sein d'un texte. Leur dimension visuelle confère à ces segments un statut particulier dans le texte. Le titre est un segment qui rompt la linéarité textuelle. Par rapport aux autres éléments textuels, on constate deux caractéristiques principales :

- une caractéristique matérielle : les titres sont généralement détachés du reste du texte : typographiquement (gras et/ou soulignés et/ou colorés et/ou numérotés) ou/et dispositionnellement (présence ou non de tabulation, sauts de lignes avant et/ou après, numérotation). Le titre porte une MFM distincte de celle qui affecte le texte lui-même ;
- une caractéristique syntaxique : la syntaxe utilisée dans les titres est généralement moins complexe que dans le reste du texte.

Le MAT fournit un cadre d'étude des titres qui permet de ne pas définir l'objet exclusivement à travers sa nature linguistique (Rebeyrolle, 2004).

Virbel (2002) propose une double fonction des titres :

- ils servent à dénoter ou à référer à l'OT titré lui-même. Dans ce cas, le titre remplit un rôle d'identificateur à la manière du nom propre.
- simultanément, ils dénotent le contenu de cet OT en en fournissant une sorte de signalétique ou de résumé significatif.

Virbel (2002) distingue les titres fonctionnels des titres thématiques.

Les ***titres fonctionnels***[7] nomment ce que constitue fonctionnellement l'OT titré (partie I, chapitre, section). Deux métaphrases différentes sont proposées :

(a) « l'OT_i (est + constitue +) un/e T_i »

(b) « l'OT_i (appartient au + relève du +)(genre + type) OT_i »

[6] Nous avons donné un exemple de fiche à la p. 38.

[7] Nous n'avons trouvé aucune occurrence de titre fonctionnel dans notre corpus.

Les ***titres thématiques*** décrivent ce dont traite l'OT titré. La métaphrase qui correspondant à ce type de titres est :

« L'OT_i (traite de + parle de + a pour (sujet + objet) + est relatif à +) OT_i »
Ce type de titre est de loin le plus fréquent dans notre corpus.

En ce qui concerne l'obsolescence, nous nous intéressons aux titres thématiques car ce sont eux qui introduisent les référents du discours et/ou les conditions temporelles, spatiales, affectives, etc. de la mise en situation d'un référent. Ainsi, dans notre corpus, nous avons constaté que la présence d'un lexique particulier entrainaît souvent la présence de segments d'obsolescence dans sa section. Ce lexique contient des termes comme « population », « monnaie », « économie », etc. Des caractéristiques temporelles comme dans le titre « Des solutions d'avenir » sont également intéressantes pour le repérage de l'obsolescence.

Jacques et Rebeyrolle (2006) distinguent deux grands types d'implication des titres thématiques dans l'organisation du contenu textuel : une *implication référentielle*, c'est-à-dire une contribution du titre à la gestion des référents du discours, et une *implication thématique*, c'est-à-dire une délimitation du thème général dans lequel s'inscrit ce dont on va parler (un domaine d'activité, un domaine de connaissances, un point de vue, une situation spatio-temporelle, etc. spécifiques). Selon ces auteurs, ces deux pôles renvoient à des processus interprétatifs différents : il s'agit dans le premier cas, d'attirer l'attention du lecteur sur un ou des référents du discours particulier(s), dans le second, de canaliser certaines de ses connaissances d'arrière-plan.

Dans une étude exploratoire sur les titres, Rebeyrolle (2004) dresse la liste des formes morpho-syntaxiques pouvant entrer dans la composition des titres thématiques[8] :

- Syntagme Nominal (SN) (l'article peut être défini pluriel, défini singulier, indéfini singulier, inexistant ou autre)

  (4.2) « La névrose d'angoisse »
- Syntagme Adjectival (SA ou SADJ) (pas d'exemple dans notre corpus)
- Syntagme Verbal (SV)

  (4.3) « Déterminer la cause »
- Syntagme Prépositionnel (SP ou SPREP)

  (4.4) « De la dépression à l'épilepsie »
- SN coordonnés

  (4.5) « Définition et aspects médico-légaux »
- SN reliés par ponctuation (virgule, deux points, point virgule, point)

  (4.6) « Le fondateur : Andrew Taylor Still »
- Titres formels[9]

  (4.7) « En bref »

[8] Tous les exemples cités ci-après sont extraits dans la mesure du possible de notre corpus [encyclo].

[9] La place de cette catégorie *titres formels* dans la classifictaion des titres thématiques est discutable : il s'agit plutôt d'un titre de type fonctionnel au même titre que « conclusion » ou « résumé ».

– Propositions subordonnées (pas d'exemple dans notre corpus)
– Phrases (interrogatives, affirmatives)

(4.8) « Le travail, c'est la santé ? »

Les titres restent toutefois des espaces infinis d'expressions, compte tenu de la variété des locuteurs, des sujets traités et d'une manière générale des combinaisons langagières possibles. Cette constatation faite, il est risqué de prétendre pouvoir proposer une grammaire réellement exhaustive de ces Objets Textuels (OT).

Ce qui nous intéresse dans la prise en compte des titres, c'est principalement la relation qu'ils mettent en œuvre entre des référents du contexte. Dans certains cas, les titres sont en relation logico-sémantique entre eux (Virbel, 2002) : les titres réfèrent à des éléments qui appartiennent à des ensembles identifiables d'objets ; ces objets entretiennent à leur tour des relations dont les titres héritent (« avant-après », « lundi-mardi-mercredi »).

Ainsi, dans le sous-corpus[ATLAS]), il est fréquent de trouver des cas comme :

(4.9) 1 « Relief »
2 « Climat »
3 « Faune et flore »
4 « Population »
5 « Economie »

La situation est très similaire avec les objets textuels suivants que nous nommons *amorces* (qui ne sont pas à proprement parler des titres mais qui se comportent comme tels) :

(4.10) – « **Nom officiel :** République d'Angola »
– « **Indépendance :** 11 novembre 1975 (du Portugal) »
– « **Superficie :** 1 246 700 km2 »
– « **Population :** 11,5 millions d'hab. (estimation 2003) »
– « **Estimation 2030 :** 16,8 millions d'hab. »
– « **Capitale :** Luanda »
– « **Langues officielles :** portugais, ombundu, kimbundu, kikongo »
– « **Monnaie :** kwanza (AOA) »

Ce sont ces récurrences de fonctionnement textuel qui nous ont conduite à créer la classe « géopolitique » dont nous avons parlé dans la section 3.5.5.

Les titres sont des objets textuels qui tiennent un rôle important dans la détermination de l'obsolescence d'un segment textuel. En règle générale, ils ne sont jamais à mettre à jour. En revanche, ils permettent de mettre en évidence, de renforcer le jugement d'obsolescence d'un segment phrastique qu'il régit. C'est parce qu'ils introduisent des élements temporels particuliers, des valeurs chiffrées ou encore l'expression de la subjectivité qu'il est possible de juger de l'obsolescence d'un segment avec plus ou moins de certitude.

Nous allons maintenant nous intéresser au cas de la position des objets dans le document.

L'importance positionnelle des objets textuels

L'organisation globale du texte est rendue explicite par l'utilisation de métaphrases qui explicitent (entre autres) l'organisation générale du texte ou encore la manière de numéroter les sections[10] :

(a) « L'auteur organise $id_0.M_0$ (texte, partie) en n parties identifiées $id_1..id_n$ »

(b) « L'auteur numérote $id_1.M_n, id_n.M_n$ »

Plus précisémment, nous nous intéressons aux objets textuels *présentatifs* et *terminatif*. Le présentatif permet de postuler au statut d'introduction et le terminatif à celui de conclusion. Les deux métaphrases associées sont :

(a) « L'auteur prélude à $id_0.M_{pc}$ (texte, partie) par un présentatif identifié id_X » ; « L'auteur confère le statut d'introduction à $id_0.presentatif$ »

(b) « L'auteur couronne à $id_0.M_{pc}$ (texte, partie) par un terminatif identifié id_X » ; « L'auteur confère le statut de conclusion à $id_0.presentatif$ »

Nous considérons les positions introductive et conclusive comme des indices potentiels pour déterminer l'obsolescence d'un segment qui viennent s'ajouter à d'autres indices comme le temps ou les valeurs chiffrées.

Nous avons remarqué que la position de conclusion (générale et de section) est souvent sujette à accueillir de l'obsolescence comme l'illustre l'exemple 4.8.

L'intégration dans l'Union européenne

§ Dans un contexte politique et social fortement marqué par la victoire de Solidarnosc, l'Église tente de consolider son influence mais la Pologne s'oriente rapidement vers une démocratie apaisée fondée sur l'alternance. C'est ainsi que les néocommunistes remportent les élections législatives de 1993 et la présidentielle de 1995 avant de céder la place à la droite libérale en 1997. Dès lors, les tensions politiques s'apaisent et les dirigeants se donnent pour but principal l'intégration de leur pays au sein de l'Union européenne. **Après l'adhésion à l'OTAN en mars 1999, les Polonais approuvent à une large majorité, en juin 2003, l'entrée de leur pays au sein de l'UE, prévue en 2004.**

Source : Corpus ATLAS (fiche Histoire - La Pologne)

Exemple 4.8 - *La typo-disposition : les paragraphes conclusifs*

Nous souhaitons également appliquer ce schéma à l'organisation interne des paragraphes pour pouvoir considérer la position des phrases dans le paragraphe et notamment les positions premières et dernières.

La métaphrase associée au paragraphe est la suivante :

« L'auteur développe un paragraphe identifié id_0 »

La métaphrase suivante permet de déterminer comment les objets textuels sont organisés à l'intérieur d'un autre objet textuel :

« L'auteur compose id_0 de $id_1, ..., id_n$ »

Associées aux métaphrases d'introduction et de conclusion présentées ci-dessus, nous pouvons dès lors considérer les premières et dernières phrases des para-

[10] id_n est un identificateur d'objet textuel ; $idi.M_j$ signifie que l'identificateur id_i doit être du type de M_j, c'est-à-dire doit appartenir à l'ensemble associé à M_j.

graphes.

Nous avons en effet constaté que les dernières phrases de paragraphe qui apparaissent également dans les derniers paragraphes de section présentent souvent une exemplification et sont souvent des segments d'obsolescence (exemple 4.9).

La voie ferrée

§ Les rails ont un double rôle : assurer le guidage des véhicules et en supporter la charge. [...] § Sur toutes les grandes lignes électrifiées, les locomotives captent le courant à l'aide d'un appareil articulé appelé « pantographe » et muni d'un archet de frottement qui glisse sous une ligne aérienne de contact, maintenue horizontale par une suspension appropriée, la caténaire. Le « retour » du courant s'effectue par les roues et les rails de roulement. Des sous-stations réparties le long de la voie ferrée reçoivent du courant à très haute tension du réseau général et fournissent le courant approprié à la caténaire. **L'électrification des réseaux continue à se développer, le plus souvent, maintenant, en courant monophasé 25 kV/50 Hz.**

Source : Corpus GLI (fiche Sciences et techniques - Les chemins de fer)

Exemple 4.9 - *La typo-disposition : les dernières phrases de paragraphes*

Les énumérations

Les métaphrases pour les énumérations sont les suivantes :

(a) « L'auteur distingue une énumération identifiée id_0 »

(b) « L'auteur agence à $id_0.M_e$ (énumération) en n points identifiés $id_1.item, ..., id_n.item$ »

(c) « L'auteur utilise le système M_{use} pour énumérer $id_1.item, ..., id_n.item$ »

Le cas des énumérations semble également intéressant pour décrire les manifestations textuelles de l'obsolescence. D'une manière générale, les énumérations (le plus souvent sous forme de listes à puce) doivent être complétées soit avec des dates, soit avec des faits plus récents. Par exemple, dans les fiches de sport, nous remarquons l'importance des résultats sportifs classés par date ou par type de championnat et sous forme d'énumération : il est nécessaire de fournir au lecteur les informations des dernières manifestations sportives en date et des derniers résultats associés.

Nous avons repéré deux situations particulièrement intéressantes d'un point de vue typo-dispositionnel dans le sous-corpus [ATLAS]. Les énumérations de résultats sportifs sont souvent suivis de « Bilan : », sorte de titre ouvrant généralement sur un segment à mettre à jour. Les énumérations de dates, de noms de personnes ou encore de noms de lieux appellent très souvent à vérifier les derniers éléments de l'énumération qu'il convient bien souvent de poursuivre et de réactualiser.

4.3.4 Conclusion

Les titres participent des mêmes métafonctions que les introducteurs de cadres :

– textuellement : des segments discursifs (les parties titrées) sont mis en évidence par la présence de titres qui ont pour fonction de regrouper des segments tels que les propositions, les paragraphes ou les cadres ;
– idéationnellement : les titres posent un critère sémantique suivant lequel les propositions suivantes sont à interpréter.

Le MAT nous intéresse spécifiquement car il met en avant le rôle d'annonce de certains éléments, la mise en saillance de certains critères sémantiques qui ont de grandes chances de se propager au-delà de l'unité dans laquelle ils apparaissent. Nous pensons particulièrement aux titres qui organisent et sont le lien entre une structure globale (le texte) et une structure locale (la phrase).

L'exemple du schéma 4.10 illustre les relations que les titres peuvent entretenir avec des élements textuels ainsi que le niveau de grain d'analyse qu'ils permettent de prendre sur un texte. Il permet de comprendre l'intérêt d'une annotation des titres pour une application visant le repérage des cadres d'interprétation (section 2.4.2 (p.57). Enfin, la prise en compte d'indices variés et de niveaux différents (en configurations d'indices) y est explicite :

– le niveau des titres ;
– des éléments lexicaux spécifiques dans les titres (lexique du temps et de l'incertitude) ;
– la position de dernier paragraphe de section ;
– des éléments temporels : syntagme nominal (« l'année 2004 »), verbes au futur et au conditionnel.

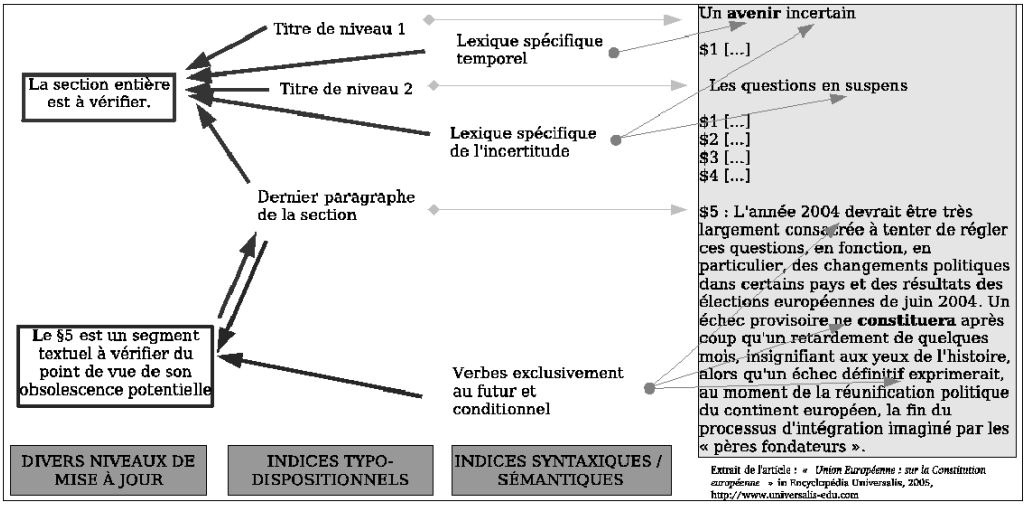


Schéma 4.10 - *Les titres pour le repérage des segments d'interprétation*

Nous avons déjà souligné le fait que les titres sont rarement des éléments qui doivent bénéficier d'une mise à jour. Il ont plutôt un rôle de prédicteurs de segments obsolescents : nous allons expliciter ce point à travers la théorie de la prédiction.

4.4 La relation de prédiction

La théorie de la prédiction (Tadros, 1985) se base sur l'idée que tout discours, qu'il soit écrit ou oral, est une interaction permanente entre deux participants. Ainsi, un *« writer »* ou *« speaker »* d'un côté, et un *« reader »* ou *« hearer »* de l'autre, communiquent à travers la médiation du texte. D'une part, le scripteur oriente le lecteur à travers le message, et de l'autre, il l'oriente à travers la réception du message. Le lecteur reconstruit alors l'organisation d'un texte par son interaction avec lui.

4.4.1 Présupposés théoriques de la prédiction

Le texte écrit est supposé interactif car au moins deux participants sont mis en jeu : le scripteur et le lecteur. Il ne s'agit certes pas d'une interaction en face à face dans la mesure où le texte écrit est non réciproque. À l'écrit, le scripteur a l'avantage de la non-interruption (par opposition avec la situation de dialogue) et il a toute liberté dans la structuration et l'organisation de son texte.

4.4.2 Définition de la prédiction

Pour Tadros (1985), la prédiction est une notion centrale lorsqu'on aborde l'étude de la structuration des textes écrits. Elle présuppose que certains signaux dans un texte ont la capacité de prédire la présence, dans la suite du discours, d'événements linguistiques particuliers.

L'auteur établit une distinction entre prédiction et anticipation :

> *« Anticipation involves guesses on the part of the reader, but prediction involves a commitment at one point in the text to the occurrence of another subsequent linguistic event. [...] Prediction is binding ; it is more in the nature of a legal contract, where predictive signals are the writer's signature confirming that he is committed to what he has said he will do. »*

Si le scripteur a posé un signal dans son texte, alors le lecteur est en mesure de prédire ce que le scripteur va faire dans la suite de son texte ; s'il n'y a pas de signal, alors le lecteur doit anticiper ce que le scripteur va faire en faisant appel à ses connaissances du monde, son bon sens (logique).

4.4.3 Le modèle

La relation de prédiction nécessite trois unités d'analyse. Tout d'abord, l'unité *paire* (*pair*) implique la présence de deux membres : le premier est prédicteur et le second est prédit, cet ordre étant invariable. Un lien est donc posé entre un item prédicteur (*predictive item*) et un item prédit (*predicted item*).

La seconde unité d'analyse est le *membre* (*member*). Un membre est constitué d'une ou plusieurs phrases (*sentence*), il peut être prédicteur ou prédit. Il est

réalisé par du texte linéaire (proposition déclarative ou sans modalité particulière) ou non (tableau ou diagramme). Concernant le membre prédicteur, ses réalisations grammaticales possibles sont les suivantes :

- proposition déclarative
  (4.11) Such goods must possess four qualities.
- proposition interrogative
  (4.12) What is wealth ?
- proposition incomplète
  (4.13) The main objections to this theory of wages are :
- proposition impérative
  (4.14) Consider Fig.29

La troisième unité d'analyse est la *phrase* (*sentence*) définie comme la réalisation linguistique minimale dans laquelle la paire ou un des deux membres de la paire est présent.

4.4.4 Les relations de prédiction

Tadros identifie six relations de prédiction. Ces six relations sont susceptibles d'apparaître conjointement dans les textes pour former des structures complexes. Ces dernières peuvent faire apparaître différentes relations en termes de discontinuité, d'emboîtement ou de chevauchement (une énumération peut être incluse dans un questionnement par exemple).

L'énumération (*Enumeration*)

La tête du membre prédicteur apporte un signal qui engage le scripteur dans une énumération. Cela annonce la présence de deux membres prédits minimum.

(a) During this period the main influences on industrial location were :
(b) This theory has been critized for the following reasons :
(c) In addition to insurance, there are a number of ways by which risk can be reduced.

Ces exemples sont des réalisations du membre prédicteur. Les membres prédits sont facilement reconnaissables car ils sont marqués de façon spécifique (italique, numérotation, ponctuation, etc.) ou par la présence d'expressions telles que « first », « secondly », « lastly », « one », « next », « then also », etc. On retrouve le fonctionnement des MIL dans la structure textuelle (*cf.* section 4.2). Ils peuvent également être marqués par la présence de verbes dans leur contexte ou de répétitions lexicales.

Acte de discours (*Advance Labelling*)

Le scripteur s'engage dans un acte de discours. Par exemple, « nous allons définir/illustrer/expliquer » l'engage à proposer une définition, à illustrer un tableau ou expliquer un diagramme.

(a) It is important, however, to distinguish between real and nominal wages.
(b) This can be illustrated by a diagram
(c) Consider Fig. 26

Compte-rendu (*Reporting*)

Le scripteur qui écrit un texte est personnellement engagé dans les opinions et idées de ce texte tant qu'il ne spécifie pas lui-même son détachement. S'il se détache d'un propos, cela prédit une explication, une évaluation future.

(a) Wages are fixed and reduced to the lowest level [...], said Quesnay (1694-1774), who first put forward this theory.
(b) Those who support the Bargaining of Wages assert that the level of Wages [...], so that, they say, [...].
(c) In their view, labour was an activ factor, [...]

Récapitulation (*Recapitulation*)

Un terme prédit le rappel d'une information déjà annoncée précédemment.

(a) As mentioned earlier, there are three types of goods.
(b) It was pointed out in the preceding section that [...]

Ce type de prédiction est très souvent combiné avec d'autres types comme l'énumération ou le compte-rendu (*reporting*).

Monde hypothétique (*Hypotheticality*)

Comme pour le *« reporting »*, cette relation est basée sur le détachement de l'auteur mais dans ce cas, ce dernier cherche à créer un monde hypothétique : il suppose et introduit une réalité hypothétique pour démontrer l'existence ou la validité de ses propos.

(a) Suppose that Dombey receives a cheque from Nickleby, [...]
(b) Consider the case of a small firm that has decided to [...]

Questionnement (*Question*)

À nouveau, l'auteur se détache de ses propos à travers le questionnement : la question posée prédit alors une réponse.

(a) What are the conditions which determine the degree of specialisation in an industry ?

Il ne s'agit là que d'une synthèse partielle de l'article de Tadros (1985). Dans cet article, l'auteur démontre que tout membre prédicteur et tout membre prédit présentent des réalisations linguistiques spécifiques qu'il est possible de lister.

4.4.5 Les titres comme prédicteurs de segments obsolescents

La relation de prédiction nous intéresse car elle permet de traiter les titres non pas comme des segments intrinsèquement obsolescents mais comme des segments

capables de prédire que les (ou certaines des) phrases qui se trouvent sous leur dépendance sont potentiellement obsolescentes ou de renforcer le caractère potentiellement obsolescent d'une phrase. En d'autres termes, il faut d'abord qu'une phrase contienne déjà certains indices pour que l'influence du titre puisse être réelle.

L'exemple de la figure 4.10 (p. 110) est explicite avec l'expression temporelle « Un avenir incertain » dans le titre. Nous avons relevé d'autres exemples dans le sous-corpus [ATLAS] qui montrent que les relations d'énumération et de questionnement peuvent, dans certains cas, permettre de prédire le phénomène recherché.

Dans l'exemple 4.11, le titre « Les SNLE, SSBN et SSGN » prédit, de par sa structure même que l'on va parler de ces trois types de sous-marins dans la section. C'est en effet ce qui se passe avec les énumérations de noms de sous-marins dans les trois paragraphes qui suivent. De plus, il s'agit de sigles, éléments que nous considérons comme des éléments relativement instables. L'ensemble de ces caractéristiques associées aux caractéristiques propres des phrases vont permettre le jugement d'obsolescence de certaines des phrases de la section. Mis à part les deux premières phrases de cet extrait, toutes les autres sont des segments d'obsolescence.

Il est intéressant de constater que d'autres mécanismes discursifs entrent en jeu dans cet extrait. Dans le premier paragraphe, un introducteur de cadre spatial va étendre sa portée sur l'ensemble des phrases du paragraphe. La spatialité est introduite dans les second et troisième paragraphes à l'aide de l'adjectif dans les syntagmes nominaux « La marine américaine » et « La situation de la flotte russe ». Enfin, des repères temporels viennent compléter cette description. Le rôle des titres s'ajoute aux rôles des autres indices.

Les SNLE, SSBN et SSGN

§ En France, les sous-marins nucléaires lanceurs d'engins sont une des pièces maîtresses de la force de dissuasion nucléaire. Ils doivent être disponibles pour exécuter une frappe en second. **Ils sont actuellement au nombre de quatre, *L'Indomptable*, *L'Inflexible*, *Le Triomphant* et *Le Téméraire* et sont armés de missiles tactiques M4 ou M45 d'une portée intermédiaire de 4 500 à 6 000 km. *Le Vigilant*, bien plus silencieux que ses prédécesseurs, doit entrer en service courant 2004 et *Le Terrible* nouvelle génération est prévu pour 2010.**

§ **La marine américaine dispose de 18 SNLE (SSBN), sous-marins américains à missile balistique (nucléaire) de type *Ohio*, équipés de 24 missiles stratégiques *Trident* d'une portée de 7 360 km. Quatre d'entre eux doivent être convertis en SSGN, sous-marins d'attaque lance-missiles, équipés de missiles de croisière *Tomahawk* pour un théâtre de guerre conventionnelle.**

§ **La situation de la flotte russe est particulière depuis l'écroulement de l'URSS. Une partie a été démantelée ou est inactive. Actuellement, la flotte russe de SNLE (porteurs de missiles de type *SS 18*, *20* et *23*) en service comprend 20 SSBN et 10 SSGN.** [...]

Source : Corpus ATLAS (fiche Sciences & Techniques - Sous-marins et bathyscaphes)

Exemple 4.11 - *Les titres prédicteurs de l'obsolescence : un exemple d'énumération*

L'exemple 4.12 montre un cas de questionnement associé à une temporalité particulière (la coïncidence). L'utilisation d'un titre renforce l'intérêt que le lecteur doit porter à la question posée et l'amène à s'interroger sur le problème posé par le rédacteur, en l'occurrence l'actualité des technologies numériques. Dans toute la section, la référence au titre est présente et l'auteur tente de répondre à la question, propose une ouverture, ouvre la discussion sur cette thématique tout en sous-entendant que cette question ne sera probablement pas résolue. Si toutes les phrases dépendantes de ce titre ne sont pas forcément à mettre à jour (nous en avons masqué quelques unes à l'aide des crochets), il est évident qu'une vérification est nécessaire en vue d'une réédition de ces informations. Parallèlement, dans la deuxième partie de l'exemple, le questionnement engage le lecteur à interpréter le reste du texte comme une vision hypothétique du futur technologique.

Des technologies nouvelles ?

§ Les technologies numériques récentes s'inscrivent dans la suite des travaux de physiciens du XIXe siècle. [...]

§ Nous atteindrons bientôt les limites de la miniaturisation avec des composants au dixième de micromètre. Les futures technologies s'appuieront peut-être sur l'électronique moléculaire, permettant d'atteindre la miniaturisation ultime, sur l'optique permettant un haut degré de parallélisme, sur les nanotechnologies enfin qui supplanteraient le silicium. [. . .]

Source : Corpus ATLAS (fiche Sciences et techniques - Les technologies numériques)

Exemple 4.12 - *Les titres prédicteurs de l'obsolescence : un exemple de question*

Le modèle de la prédiction textuelle développe l'idée qu'un texte est un objet interactif dans lequel le scripteur utilise des signaux spécifiques pour guider ses lecteurs vers telle ou telle interprétation. Nous avons vu, à travers le modèle du MAT que les éléments de typo-disposition pouvaient également jouer ce rôle. Dans ces deux cas, les objets textuels manipulés sont des guides interprétatifs capables d'indiquer l'obsolescence d'un segment.

4.5 Conclusion

Dans ce chapitre, nous avons posé les bases théoriques de notre analyse linguistique. Nous avons d'abord précisé l'orientation générale de notre point du vue sur les textes en nous inscrivant dans une approche fonctionnelle du discours. Nous en avons décrit les concepts inhérents et nécessaires.

Nous avons ensuite suggéré l'idée qu'il existe des indices linguistiques posés de manière intentionnelle ou non par le scripteur qui influent sur l'interprétation de certains segments textuels et amènent à les considérer comme potentiellement obsolescents. Nous pensons que ces indices textuels et discursifs, s'ils sont envisagés en termes de configurations d'indices peuvent agir comme des marqueurs à part entière. Ils permettraient alors d'identifier les segments obsolescents.

Nous avons montré en quoi l'approche énonciative des discours pouvait nous être utile pour décrire le phénomène de l'obsolescence.

Trois modèles/hypothèses ont finalement été décrits. Tout d'abord, l'encadrement du discours nous permet d'envisager une segmentation à la fois textuelle et idéationnelle. Puis le modèle de l'architecture textuelle nous permet de considérer les titres ainsi que la position des unités textuelles dans les documents comme des unités linguistiques à part entière. Enfin, le modèle de la prédiction justifie la prise en considération des titres comme des prédicteurs potentiels de l'obsolescence.

Bilan de la seconde partie

Cette partie est consacrée à la description linguistique des segments d'obsolescence.

Nous avons d'abord présenté les expressions linguistiques apparaissant fréquemment dans ces segments. Cette présentation est l'aboutissement d'une observation manuelle minutieuse des segments d'obsolescence tels qu'ils ont été annotés manuellement par les experts (*cf.* chapitre 2.3).

Les aspects temporels sont centraux lorsqu'on s'intéresse à l'obsolescence. C'est en effet sur eux que le découpage des événements et la relation à la situation d'énonciation est d'abord perçue. Nous avons également montré que le temps n'est pas le seul aspect à prendre en compte : les expressions aspectuelles et modales se sont révélés des pistes de recherche intéressantes. Le point de vue du rédacteur, souvent exprimé implicitement, semble également un critère pertinent pour évaluer la validité d'une information. Enfin, la très large et hétérogène catégorie des entités nommées (valeurs chiffrées, lieux, noms propres, lexiques spécifiques, etc.) regroupe des expressions linguistiques qui semblent productives pour l'obsolescence.

Aucun de ces indices linguistiques n'est capable de rendre compte de l'obsolescence s'ils est considéré indépendamment des autres. Aussi est-il fondamental de les envisager en combinaisons. Dans le tableau suivant (4.1), est répertorié l'ensemble des indices susceptibles de participer à ces configurations d'indices pour la description et le repérage des segments d'obsolescence.

Catégorie générique	Typage 1	Typage 2	Réalisations
Temps	* anaphorique * déictique	* antériorité * postériorité * coïncidence * indéterminé	– temps des verbes – adverbiaux – syntagmes nominaux – dates
Aspect	* inaccompli		– temps verbaux – adverbiaux – périphrases verbales
suite du tableau page suivante...			

Catégorie générique	Typage 1	Typage 2	Réalisations
	* imperfectif		– syntagmes verbaux
	* sécant		– temps verbaux – adverbiaux
	* inchoatif		– périphrases verbales – négation d'un terminatif
	* itératif		– adverbiaux
	* progressif		– locutions prépositionnelles
Modalité	* d'énoncation	* exclamation * interrogation * assertion	– ponctuation
	* d'énoncé	* évidentialité	– syntagmes prépositionnels
		* commentaire	– structures impersonnelles
		* argumentation logique	– argumentatifs
		* subjectivité	– adverbes affectifs
Entités nommées	* Lieux		– syntagmes prépositionnels – syntagmes nominaux – noms propres

suite du tableau page suivante...

Catégorie générique	Typage 1	Typage 2	Réalisations
	* Personnalités physiques et morales		– noms et noms propres – sigles – noms de marques, abréviations – syntagmes nominaux
	* Mesures		– chiffres – abréviations, noms de mesure
	* Superlatifs		– syntagmes nominaux
	* Géopolitique		– noms – syntagmes nominaux – sigles
Position dans la phrase			– début de phrase – fin de phrase – position « amorce »
Position de la phrase			– début de paragraphe – fin de paragraphe
Position du paragraphe			– début de section (introduction) – fin de section (conclusion) – + niveau de la section
Position dans un titre			– présence – absence – + niveau du titre
			Fin du tableau

TAB. 4.1 - *Résumé des indices susceptibles d'être pertinents pour l'obsolescence*

Parce qu'il n'y a pas un seul indice linguistique propre à l'obsolescence, la prise en considération d'une grande variété d'indices de types différents, aux sémantismes variés et relavant de niveaux structurels distincts nous semble fondamental pour aborder le repérage de l'obsolescence en termes de faisceaux d'indices. Cette

démarche, orientée *multi-indices* et variabilité du grain d'analyse, nous a amené à considérer l'exploitation de modèles discursifs spécifiques, l'encadrement du discours et le modèle d'architecture textuelle, de manière centrale. L'ensemble des éléments linguistiques décrits dans cette partie sont des éléments tangibles venant conforter la validité des modèles discursifs présentés.

C'est ainsi que les niveaux discursifs suivants peuvent être pleinement exploités :

- le niveau **intra-phrastique** correspond aux expressions linguistiques internes à la phrase : adverbiaux temporels, temps verbaux, entités nommées, adverbes exprimant la position du rédacteur face à ses dires, expressions à valeur aspectuelle, etc.
- le niveau **positionnel phrastique** rend compte de la position des expressions linguistiques au sein de l'unité phrase. Il permet notamment d'étudier le caractère potentiellement cadratif (position initiale de proposition) ou non (position finale) de certaines unités. C'est principalement à travers l'hypothèse de l'encadrement du discours que la position des indices dans la phrase est envisagée.
- le niveau **positionnel textuel** rend compte de la position d'une unité au sein d'une unité plus large : par exemple l'unité « phrase » au sein de l'unité « paragraphe » (première phrase ou dernière phrase du paragraphe, phrase seule dans un paragraphe) ou l'unité « paragraphe » au sein de l'unité « document » (premier paragraphe ou dernier paragraphe du document ou de la section) ou encore l'unité « phrase » au sein de l'unité « section » (à travers les niveaux des titres notamment). Le modèle de l'architecture textuelle est exploité dans cette optique.
- le niveau **hiérarchique** correspond au fait qu'une unité (par exemple une phrase) est sous la dépendance d'une autre unité (par exemple un titre) mais sans véritablement être incluse. C'est une relation qui s'apparente à la rection. Un titre est susceptible de contenir des indices de type intra-phrastique qui vont potentiellement influencer le caractère obsolescent ou non des unités sous sa dépendance. Le modèle de l'architecture textuelle est ici exploité ainsi que la théorie de la prédiction.
- le niveau **externe** qui fait référence au type de document ou au domaine au sein duquel un texte est rédigé.

Notre objectif est maintenant de comprendre les mécanismes organisationnels des éléments susceptibles d'apparaître à n'importe quel niveau discursif. Quelles combinaisons sont pertinentes pour l'obsolescence ? Nous envisageons les combinaisons d'indices non pas en termes de contexte pertinent ou non (*cf.* l'exploration contextuelle, p. 27) mais plutôt en termes d'associations à l'image des « stéréotypes organisationnels » de Widlöcher (2008).

Nous allons maintenant décrire le dispositif expérimental que nous avons élaboré dans le double objectif de décrire quantitativement et qualitativement les segments d'obsolescence et de répondre au besoin applicatif concret (le prototype d'aide à la mise à jour automatique de l'obsolescence).

Troisième partie

Silence, on tourne !

Cette troisième partie présente le dispositif expérimental mis en place pour rendre compte du phénomène de l'obsolescence dans les textes encyclopédiques. Le terme *« dispositif expérimental »* est pris au sens de Habert (2005) qui le définit comme un *« montage d'instruments, d'outils et de ressources servant à produire des « faits » dont la reproductibilité et le statut (l'interprétation) font l'objet de controverses »*.

Nous proposons de nommer la méthodologie que nous avons mise en place dans ce travail : la méthode RIO (**R**epérage de l'**I**nformation **O**bsolescente). Elle permet de rendre compte à la fois des outils, ressources et instruments utilisés et implémentés et de leur montage, de leur organisation et leur interdépendance. Nous avons cherché à rendre cette méthodologie reproductible et adaptable pour d'autres tâches. Le schéma 4.13 présente les trois étapes principales de la méthodologie.

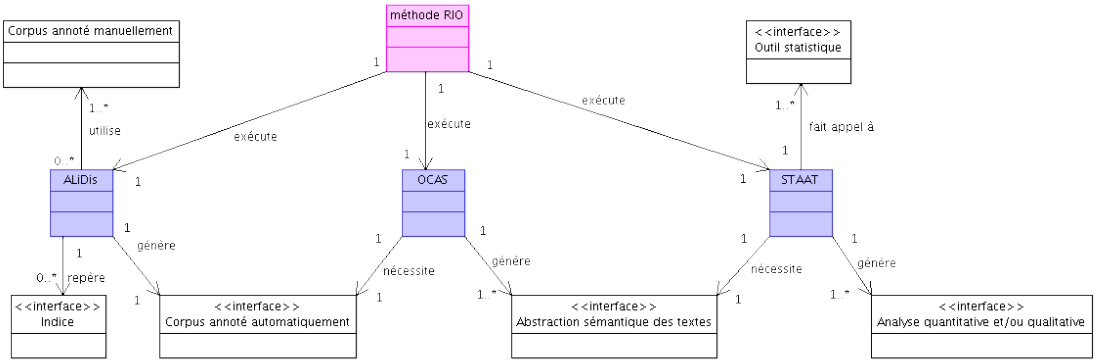


Schéma 4.13 - *La méthode mise en place (méthode RIO)*

La première étape consiste à repérer dans notre corpus annoté manuellement ([ENCYCLO]), les expressions linguistiques susceptibles d'être de bons indices de l'obsolescence (*cf.* les indices linguistiques et discursifs décrits dans la partie II) : c'est le rôle de l'outil ALIDIS (**A**nnotation **LI**nguistique des **DIS**cours). Il produit en sortie un corpus annoté automatiquement des indices potentiels de l'obsolescence.

À partir des résultats fournis par ALIDIS, l'étape 2 consiste à transformer le format des données textuelles (en XML) en un format matriciel afin de permettre un traitement statistique des indices linguistiques et discursifs. Cette transformation est basée sur un modèle pivot (*i.e.* un modèle conceptuel des données) : c'est l'objectif de l'outil OCAS (**O**util de **C**réation d'**A**bstraction **S**émantique).

Enfin, les données générés par OCAS font leur entrée dans l'outil STAAT qui produit une analyse statistique. Cette analyse statistique se fait en trois étapes : un module de statistiques descriptives basiques, une Analyse en Composantes Principales (ACP) et enfin un module d'apprentissage automatique (AA). L'intérêt est double :

(i) décrire qualitativement et quantitativement les indices potentiels de l'obso-

lescence ainsi que les segments annotés obsolètes ;

(ii) apprendre automatiquement des règles de combinaisons d'indices pour le repérage automatique de l'obsolescence dans les textes encyclopédiques.

Cette partie est divisée en cinq chapitres. Les chapitres 5, 6 et 7 présentent et décrivent chacune des étapes de la méthode RIO[11]. Le chapitre 8 propose une discussion des résultats ainsi qu'une mise en perspective avec des travaux de T.A.L. qui sont proches. Enfin le chapitre 9 met en place la procédure d'évaluation du prototype d'aide à la mise à jour de textes encyclopédiques.

[11] Les schémas de cette partie ont été crées en UML (*Unified Modelling Language*, Langage Unifié pour la Modélisation). Nous rappelons dans l'annexe A (p. 249) les principaux objets utilisés. Pour plus d'informations, nous renvoyons le lecteur aux normes de ce standard (`www.uml.org`).

Chapitre 5

Étape 1 : Outil ALIDIS (Annotation LInguistique des DIScours)

Dans ce chapitre, nous présentons l'outil ALIDIS (***A**nnotation **LI**nguistique des **DIS**cours*) développé pour repérer et annoter automatiquement les indices linguistiques et discursifs potentiellement pertinents pour la délimitation automatique des segments d'obsolescence.

Les différents modules de ALIDIS ont été construits de manière à rester le plus possible en adéquation avec les théories et hypothèses linguistiques exposées dans les chapitres 3 (p. 65) et 4 (p. 89). Ainsi, nous proposons l'étude de quatre analyseurs sémantiques et deux analyseurs structurels (*cf.* schéma 5.1, p. 126) :

1. le traitement du temps ;
2. le traitement de l'aspect ;
3. le traitement des entités nommées (lieux, personnes, chiffres et superlatifs) ;
4. le traitement de la modalité ;
5. le traitement de la position des indices dans la phrase ;
6. le traitement de la typo-disposition.

Le schéma 5.1 récapitule ces objectifs.

Ancré dans le domaine du T.A.L., ce travail d'implémentation exploite naturellement un certain nombre de techniques présentées dans le chapitre 1 (p. 15). Nous présentons d'abord la plateforme LINGUASTREAM qui nous a permis de développer nos analyseurs.

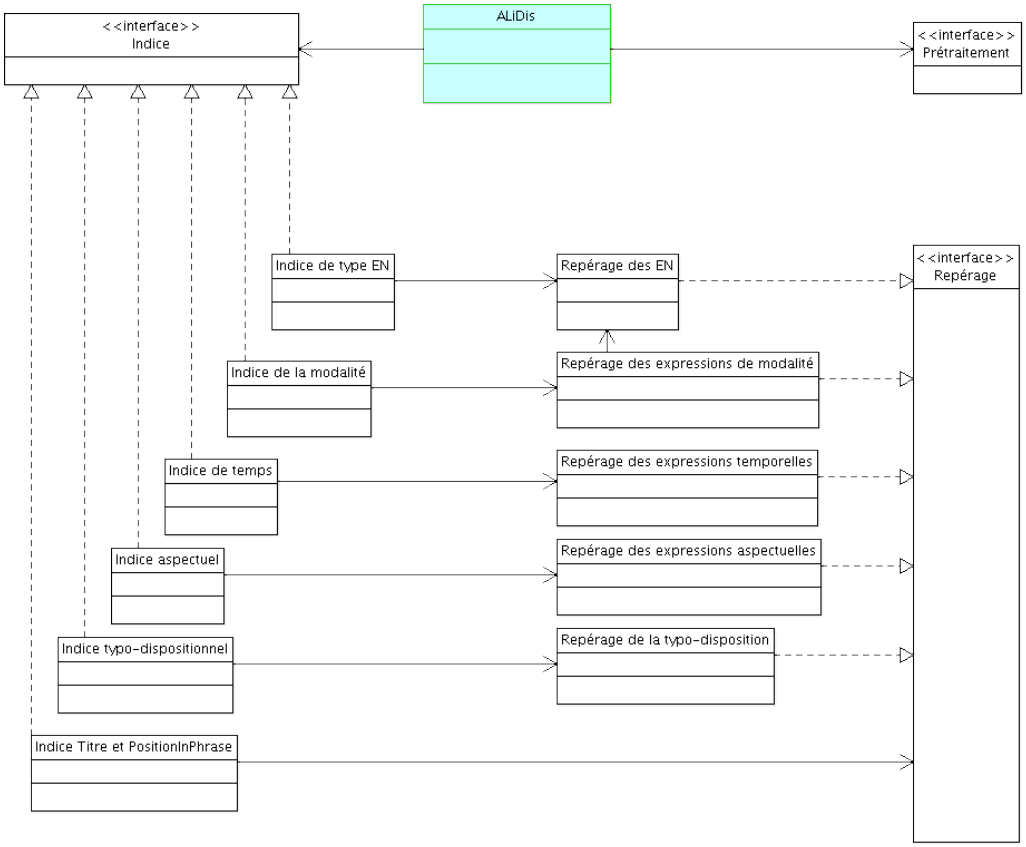


Schéma 5.1 - *L'outil* ALIDIS

5.1 LINGUASTREAM, une plateforme d'expérimentation pour le T.A.L.

Nous avons utilisé LINGUASTREAM[1], plateforme dédiée au traitement du langage naturel développée au GREYC (Caen) par F. BILHAUT et A. WIDLÖCHER (Widlöcher et Bilhaut, 2005).

LINGUASTREAM est fondée sur la notion de *chaîne de traitement* : envisager un traitement sous forme de chaîne consiste à décomposer un problème (linguistique), souvent complexe, en plusieurs sous-tâches bien délimitées, connues et relativement consensuelles dans le domaine et donc plus facilement automatisables. Dans LINGUASTREAM, ces tâches localisées prennent l'apparence de *modules* de traitements *a priori* distincts (*i.e.* qui nécessitent souvent des formalismes différents), connectés entre eux et qui collaborent de manière transparente pour l'utilisateur.

Certaines techniques de T.A.L. comme l'étiquetage morpho-syntaxique ou l'analyse syntaxique sont aujourd'hui suffisamment abouties pour fournir des résultats suffisamment fiables. La plateforme LINGUASTREAM permet de tirer parti de ces résultats afin d'élaborer des systèmes réalisant des tâches de plus haut niveau.

LINGUASTREAM permet également de revenir aisément à la fois sur les résultats des repérages et annotations automatiques et sur les programmes eux-mêmes : cette navigation entre les données initiales, les programmes et les résultats autorise des modifications des traitements effectuées de manière incrémentale. Nous rejoignons le constat de Habert (2005, p. 115) qui insiste particulièrement sur la tendance actuelle à envisager l'annotation de corpus comme un processus itératif et perpétuel : le corpus n'est plus constitué de manière stable et immuable, il peut au contraire bénéficier d'annotations de niveaux différents sur lesquelles il est facile de revenir et qu'il est aisé de modifier en fonction de ses propres objectifs. Toujours selon Habert (2005), ce type d'annotation *en flux* nécessite de savoir combiner instruments et ressources. Cette combinaison s'effectue par une *trousse à outils* qui allie langages de scripts, bases de données, tableurs, etc. De plus, LINGUASTREAM facilite le repérage et l'annotation d'indices linguistiques à granularité variable (mot, syntagme, expression, position textuelle, position des phrases et des paragraphes, titres, etc.) et de types linguistiques diversifiés (temps, modalité, aspect, noms propres, etc.). Il est donc intéressant de disposer de formalismes différents et adaptés au type d'expressions recherchées. Nous n'exploitons pas l'ensemble des possibilités de la plateforme : nous nous limitons à l'utilisation des expressions régulières et macro-expressions régulières, des lexiques ou encore des grammaires ProLog.

Il ne s'agit pas de procéder uniquement à un repérage de surface des indices. À chacune des expressions repérées, un typage sémantique ciblé pour la recherche de l'obsolescence est associé. Ce processus d'abandon progressif des formes de surface permet de s'abstraire de la linéarité des textes : au final, ce n'est pas un

[1] `http://www.linguastream.org`

texte qui est traité, ni des suites de mots mais des représentations sémantiques et abstraites particulières (*cf.* chapitre 6, p. 161).

Le repérage des indices n'est pas la finalité de ce travail. Aussi, la possibilité d'exporter facilement les résultats vers d'autres outils et d'autres formats constitue un élément important dans LINGUASTREAM.

La chaîne de traitement construite avec LINGUASTREAM a pour objectif le repérage des indices susceptibles d'être de bons indices de l'obsolescence. Un des intérêts majeurs d'une chaîne de traitement est la modularité et la décomposition des objectifs en sous-tâches spécialisées relevant très souvent de niveaux d'analyse variés : nous exploitons principalement les lexiques, les grammaires ProLog et les macro-expressions régulières (MRE). La figure 5.2 présente la chaîne de traitement créée dans l'environnement LINGUASTREAM.

Nous présentons d'abord les prétraitements effectués sur le corpus (segmentation en mots, étiquetage morpho-syntaxique, segmentation en phrases et création de ressources lexicales, *i.e.* les prétraitements), puis les programmes de repérage des indices linguistiques : le traitement du temps, de l'aspect, de la modalité, des entités nommées, de la position des indices et de la typo-disposition.

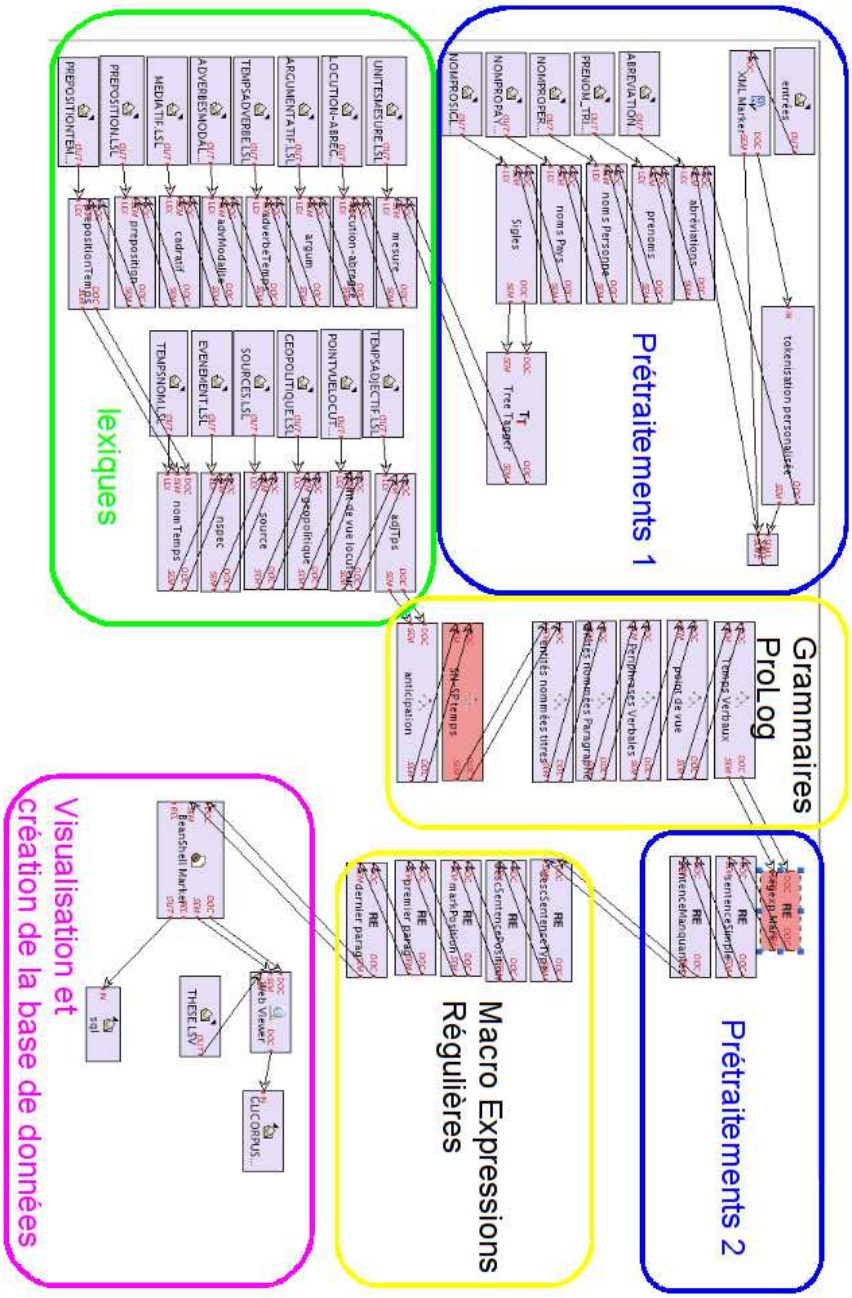


FIG. 5.2 - *Chaîne de traitement pour le repérage d'indices linguistiques (*ALIDIS*)*

5.2 Traitements de base pour une analyse en T.A.L.

Un certain nombre de traitements de base sont indispensables lorsqu'on traite automatiquement la langue. Ainsi, comme le montre le schéma 5.3 (p. 130), notre analyse nécessite une étape de segmentation des textes en mots. Dans notre cas, nous exploitons également une structure XML spécifique qui rend compte notamment du découpage des titres, des paragraphes ou encore qui nous informent sur le type de texte (par rubrique notamment : Géographie, Histoire, Sciences et techniques, etc.). Le module de délimitation des phrases du textes est basé sur cette architecture XML. L'exploitation de lexiques spécifiques constitue également une étape importante de prétraitements. Enfin, nous projetons une analyse morpho-syntaxique sur l'ensemble des mots délimités (utilisation de l'étiqueteur morpho-syntaxique TreeTagger).

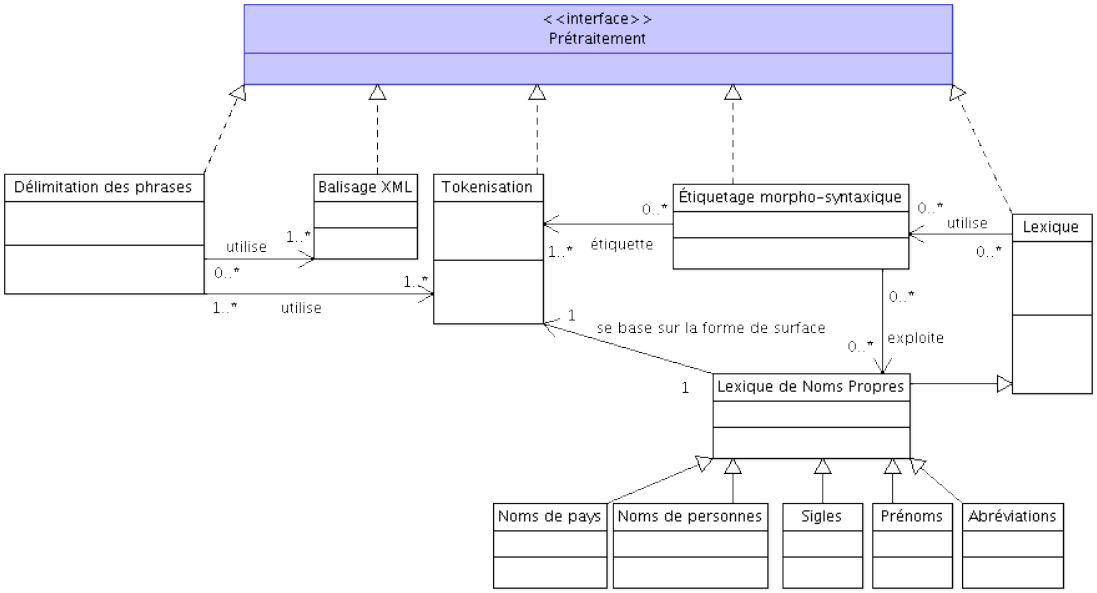


Schéma 5.3 - *Les prétraitements nécessaires pour nos analyseurs*

5.2.1 Segmentation : découpage du texte en mots

Le découpage en mots est une étape triviale mais cruciale dans un système de T.A.L. : un effet de réaction en chaîne est fréquent, un choix de découpage initial pouvant avoir des répercussions sur de nombreuses étapes ultérieures de la chaîne de traitement. C'est une tâche relativement bien définie et de nombreux outils existent. Ils sont souvent dédiés à une tâche précise. Nous avons ainsi constaté que les choix de découpage des mots faits par certains segmenteurs ne sont pas adaptés à notre problématique et qu'ils entraînent un certain nombre de problèmes délicats à gérer *a posteriori* pour le repérage des indices linguistiques.

Nous avons créé un programme de tokenisation qui nous permet de contrôler les effets de certains découpages.

Notre principale difficulté concerne le traitement des abréviations et des sigles, expressions très souvent composées de points. Dans la plupart des cas, l'étiqueteur morpho-syntaxique que nous utilisons, TreeTagger, dans sa phase de segmentation, exclut le point dans le découpage des mots. Nous souhaitons pour notre part que, dans le cas des abréviations et des sigles notamment, le point fasse partie du mot, ceci afin de faciliter notamment le repérage ultérieur des phrases.

Les sigles sont donc repérés dans un premier temps par une expression régulière basique qui ne prend en compte que les suites de majuscules éventuellement intercalées par une espace, un point ou un tiret. Nous disposons également d'un lexique de sigles constitué à partir des dictionnaires et encyclopédies des Éditions Larousse[2].

Concernant les abréviations (« etc. », « hab. », « Mr »), nous avons fait le choix de les traiter à l'aide de lexiques constitués à la main et à partir des bases encyclopédiques Larousse[3].

Les autres mots sont délimités à l'aide d'un système par expressions régulières relativement simple qui considère un mot comme une suite de caractères situés entre deux espaces ou entre une espace et un signe de ponctuation.

Une seconde raison nous a amenée à créer notre propre module de segmentation en mots : l'outil d'étiquetage morpho-syntaxique (le TreeTagger) ne propose pas toujours des étiquetages adaptés à nos besoins. Ainsi, par exemple, pour le repérage des expressions temporelles, il nous faut distinguer les adjectifs ordinaux des adjectifs cardinaux.

Le module d'expressions régulières de LINGUASTREAM autorise l'association d'annotations sémantiques spécifiques en fonction des patrons reconnus. Par ailleurs, TreeTagger accepte des pré-annotations morpho-syntaxiques. Nous avons ainsi forcé l'étiquetage morpho-syntaxique du TreeTagger de certaines expressions. Il s'agit notamment :

– des adjectifs ordinaux (« 3ème », « 2e ») et cardinaux (« 16 », « 168 000 ») ;

[2] À l'aide de transformations XSLT, nous avons extrait les expressions XML renvoyant à la notion de sigle. Étant donné que dans la base Larousse nous disposons également de la forme complète du sigle, nous considérons comme un sigle à la fois l'acronyme et sa version complète.

[3] À l'aide de transformations XSLT également.

- des noms des pages html et adresses mails ;
- de certains symboles particuliers comme ® ;
- des prénoms, de noms de personnalités et de noms de pays (également créés à partir des bases encyclopédiques Larousse) ;
- et évidemment des sigles et des abréviations.

5.2.2 Étiquetage morpho-syntaxique

Pour l'étiquetage morpho-syntaxique, nous utilisons l'outil TreeTagger. Dans la plateforme LINGUASTREAM, il est associé à un module externe.

TreeTagger est développé à l'*Institute for Computational Linguistics of the University of Stuttgart*[4]. Il effectue par défaut une segmentation[5], une lemmatisation et un marquage morpho-syntaxique des unités lexicales. Il associe à chaque mot les traits suivants : un lemme (`lemma`), une étiquette morphologique (`tag`) et une sous-étiquette morphologique (`stag`).

TreeTagger est un outil fiable et robuste, libre et gratuit pour la recherche et surtout, il est modulable. Comme nous l'avons vu dans la section précédente, il est suffisamment souple pour accepter des pré-annotations morpho-syntaxiques.

5.2.3 Segmentation en phrases

La segmentation en phrases est centrale dans notre travail pour deux raisons. Tout d'abord, elle est nécessaire pour situer les indices selon leur position à l'initiale ou en finale de phrase : nous considérons la position des indices comme une information forte (*cf.* section 4.2, p. 98).

De plus, la phrase est le niveau de segmentation minimal de réalisation de l'obsolescence : nous avons en effet contraint les annotations manuelles des experts à considérer cette unité comme la taille de segment d'obsolescence minimal (*cf.* section 2.3.2, p. 46).

Le repérage des phrases est donc pour nous une étape essentielle. Nous avons déjà mentionné le prétraitement des sigles et des abréviations. Un autre prétraitement pour le repérage des phrases est également nécessaire : il s'agit de la prise en compte des éléments entre parenthèses.

Une fois ces deux prétraitements appliqués, nous considérons une phrase comme une suite de mots situées entre une marque de début de paragraphe ou une balise de fin de phrase et un signe de ponctuation fort (« ? », « ! », « ... ») ou une balise de fin de paragraphe[6]. Les deux-points et le point-virgule ne sont pas considérés comme des ponctuations fortes et ne permettent donc pas de délimiter nos unités phrases[7].

[4] `http://www.ims.uni-stuttgart.de/projekte/corplex/Treetagger/`

[5] Que nous n'exploitons pas.

[6] Nous avons rencontré des cas où la phrase ne contient pas de ponctuation finale.

[7] Ce choix est lié au fait que nous considérons la phrase comme l'unité minimale pour un segment d'obsolescence (*cf.* section 2.3, p. 42).

5.2.4 Utiliser des ressources : constitution de lexiques

La majorité des applications T.A.L. fonctionnent en combinant des règles (des « grammaires ») avec des bases de connaissances (des « lexiques »). C'est notamment le cas pour les systèmes de correction orthographique, d'analyse syntaxique, ou encore d'extraction de connaissances. Nous n'échappons pas à cette règle et avons développé un certain nombre de ressources lexicales plus ou moins ciblées vers notre objectif.

Démarche générale

Plusieurs raisons nous ont amenée à créer nos propres ressources lexicales. Tout d'abord, il s'avère difficile de trouver des ressources disponibles pour le français. De nombreux auteurs mentionnent la possibilité d'utiliser les ressources qu'ils ont constituées mais dans la plupart des cas, la démarche d'acquisition de ces lexiques est contraignante[8].

Parmi les ressources disponibles pour le français dans la communauté T.A.L., citons par exemple les tables du lexique-grammaire[9]. Dans ce cas, les ressources sont exclusivement utilisables *via* les environnements INTEX et UNITEX.

À l'opposé, la plateforme GATE par exemple, dans laquelle les ressources textuelles sont entièrement disponibles et utilisables indépendamment de la plateforme, rend compte d'un modèle que nous jugeons pertinent en termes d'échanges de connaissances et de ressources. Nous espérons y contribuer en mettant l'ensemble des ressources créées dans le cadre de ce travail à disposition[10].

La seconde raison qui nous a amenée à créer nos propres ressources est que, lorsque la ressource existe, elle est payante. Ainsi l'association ELRA qui a pour but de promouvoir la production de ressources langagières[11] propose un certain nombre de ressources intéressantes mais payantes. Et pour accéder à de telles ressources, il faut disposer d'un budget dédié, ce qui n'est pas le cas pour ce travail.

Les lexiques créés ne sont pas spécialement innovants (lexiques de noms de mesures, d'abréviations, de noms, adjectifs, adverbes de temps, etc.). Mis à part le côté laborieux de leur constitution (recherche dans nos corpus, dans des dictionnaires papiers, dans les bases Larousse[12], dans des dictionnaires en ligne, dans des grammaires du français ou encore dans des articles de recherche), leur élaboration a été finalement relativement rapide.

---

[8] Souvent, il faut contacter directement les auteurs ou alors utiliser l'outil, le système informatique qui a été développé conjointement.

[9] La description fine d'un mot consiste à examiner son comportement pour le maximum de propriétés syntaxiques pertinentes pour sa catégorie grammaticale. Ressources disponibles sur : `http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Presentation.html`.

[10] Les ressources créées dans le cadre de ce doctorat sont disponibles sous licence LGPL, librement téléchargeables sur `http://marion.laignelet.free.fr`.

[11] *European Language Resources Association*, `http://catalog.elra.info/`

[12] Nous avons également exploité la base encyclopédique Larousse pour créer notamment les lexiques de « noms propres » (pays, personne, prénoms, sigles et abréviations).

Les lexiques constitués sont des listes structurées et enrichies d'informations syntaxiques et sémantiques. Pour chaque lexique, nous définissons des annotations sémantiques particulières sous forme de structures de traits qui s'appliquent à chaque terme du lexique. Ces informations (syntaxiques et sémantiques) ont pour rôle de contraindre l'utilisation des entrées lexicales à des patrons syntaxiques déterminés. Concernant les informations de type sémantique, elles sont primordiales dans la construction de la représentation abstraite des corpus (*cf.* outil OCAS, chapitre 6, p. 161).

Parallèlement aux lexiques *classiques*, des lexiques adaptés à notre problème sont nécessaires.

Description des lexiques créés

Il y a dix-neuf lexiques. Cinq d'entre eux (lexiques d'abréviations, de prénoms, de noms de personnes, de noms de pays et de sigles) ont déjà été décrits au moment de la présentation du programme de tokenisation (*cf.* p. 131). La taille des lexiques est variable : de quelques dizaines d'entrées pour un lexique comme celui des prépositions, à plusieurs milliers comme le lexique des noms de personnes.

Les autres lexiques servent à repérer et annoter des expressions comme :

- des unités de mesure (« euros », « kg »)
- des titres distinctifs (« Madame », « Monseigneur », « Saint », « Mr »)
- des éléments argumentatifs (« par exemple », « mais », « enfin », « premièrement »)
- des adverbes de temps (« encore », « prochainement »)
- des adverbes exprimant un jugement subjectif du rédacteur (« malheureusement », « hélas », « peut-être »)
- des noms de temps (« année », « période », « jour »)
- des prépositions génériques (« de », « à »)
- des prépositions de temps (« depuis », « dans le courant de »)
- des adjectifs de temps (« récent », « actuel », « ancien »)
- des adverbes *médiatifs* (« selon », « à propos de », « quant à »)
- des unités lexicales permettant le repérage d'expressions exprimant le point de vue du locuteur (« impression », « prévision », « prévoir », « estimer », « croire »)
- des listes de sources d'information (« rapport », « débat », « recherche »)
- des expressions propres au domaine de la géopolitique (« densité de population », « démographie », « monnaies »)
- des expressions référant à des événements précis de l'histoire (« Ve République », « Moyen-âge »)

Ces lexiques ont été construits sur la base de nos corpus, puis enrichis à l'aide de dictionnaires papiers et électroniques. Le dictionnaire des synonymes développé par le CRISCO[13] a constitué une base de référence importante dans la constitution

[13] http ://www.crisco.unicaen.fr/cgi-bin/cherches.cgi

de ces lexiques. Enfin, nous avons exploité les travaux en linguistique qui mettent leurs ressources à disposition : les travaux d'A. Borillo nous ont été particulièrement utiles pour tout ce qui concerne le traitement du temps (Borillo, 2003a,b, 1998a, 1983) et celui de l'espace (Borillo, 1998b, 1990, 1988).

La figure 5.4 montre un lexique créé à l'aide de l'éditeur de la plateforme LINGUASTREAM (il s'agit du lexique de noms de mesure). Dans la fenêtre de gauche sont listés tous les mots du lexique. La fenêtre en haut à droite permet de fournir les différentes formes possibles pour une même entrée : dans l'exemple, l'entrée principale est « euro », il peut être au pluriel, on a donc la forme « euros »[14] et enfin le signe « € ». La fenêtre en bas à droite permet d'associer à chacune des entrées du lexique un certain nombre d'annotations. Nous avons déjà mentionné le fait que ces annotations peuvent être de type syntaxique ou de type sémantique. Dans cet exemple, les annotations sont essentiellement sémantiques : elles renseignent sur le type de mesure (en l'occurrence une monnaie) et sur sa capacité à évoluer ou non dans le temps (*i.e.* à accompagner une valeur peu stable dans le temps, qui évolue beaucoup).

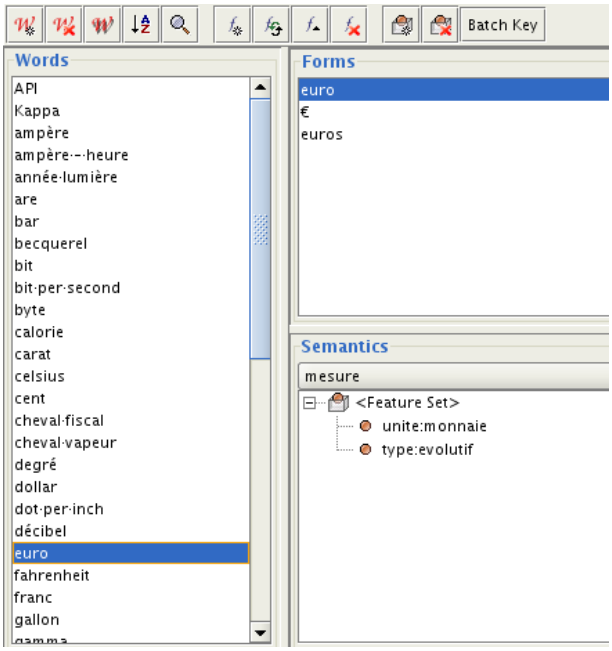


FIG. 5.4 - *Créer des lexiques avec le LexiqueMarker de* LINGUASTREAM

Une des limites de nos lexiques, concerne l'homonymie. Par exemple, le mot « car » est une entrée du lexique de mots argumentatifs. Le composant LEXICONMARKER de LINGUASTREAM ne permet pas de sélection sur la nature des mots

[14] Il n'est normalement pas nécessaire de lister les différentes formes fléchies pour un même mot mais nous avons rencontré certains bugs de la plateforme qui nous ont amenée à faire le choix de tout mettre. La version actuelle de LINGUASTREAM corrige ce bug et permet de traiter soit le lemme (fourni par TreeTagger) soit la forme de surface.

(nom, verbe, etc.). Toutes les occurrences de « car » sont donc annotées comme un argumentatif même dans les cas où il s'agit d'un nom. Nous résolvons partiellement ce problème en n'exploitant aucun lexique de manière brute : ils sont tous réutilisés dans des grammaires locales (écrites en ProLog) qui jouent en quelque sorte le rôle de filtres syntaxiques et/ou contextuels.

Nous allons maintenant présenter les programmes de repérage et d'annotation des expressions susceptibles de marquer l'obsolescence : traitement du temps, des valeurs aspectuelles, de l'expression de la position du rédacteur face à ses propos, des entités nommées, des éléments de la typo-disposition et de la position des indices dans la phrase. Pour chacun des analyseurs, nous décrivons leur mise en œuvre globale et nous proposons une évaluation qualitative (analyse d'exemples) et quantitative (selon les mesures de précision et de rappel (décrites à la page 18). Ces évaluations ont été menées sur environ $1/10^e$ du nombre de phrases du corpus (sous-corpus ATLAS et sous-corpus GUL).

5.3 Le traitement automatique du temps

Les aspects temporels, comme nous l'avons vu dans le chapitre 3 (p. 65), prennent des formes très variées dans les textes. Nous traitons les deux grandes classes que sont les temps verbaux d'un côté, les adverbiaux temporels de l'autre.

Le schéma 5.5 rend compte des différents modules utiles au repérage et à l'annotation sémantique des expressions de temps. La grammaire dédiée au repérage des expressions temporelles est contrainte par deux types de prétraitements : l'étiquetage morpho-syntaxique et l'utilisation de lexiques spécifiques (lexique d'adverbes de temps, de prépositions génériques, de prépositions temporelles, de noms et d'adjectifs de temps, de noms d'événements historiques.

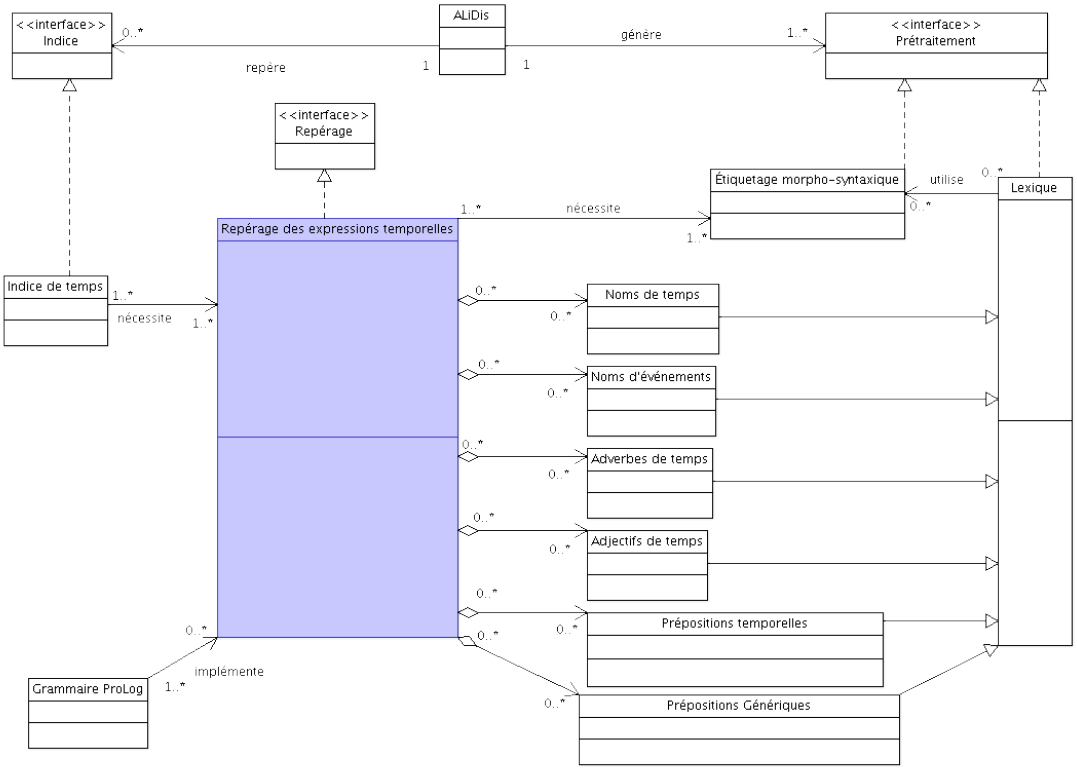


Schéma 5.5 - *Le traitement des expressions de temps*

5.3.1 Les temps verbaux

Pour les temps verbaux, nous nous sommes basés sur l'étiquetage du TreeTagger pour les formes simples. Pour les formes composées, une grammaire ProLog repère les verbes au passé composé en se basant sur la présence d'un auxiliaire (« être » ou « avoir ») suivi d'un verbe au participe passé. La présence d'un adverbe entre l'auxiliaire et le participe est autorisée.

5.3.2 Les adverbiaux temporels

Le traitement des adverbiaux temporels est plus complexe que celui des temps verbaux :

- il regroupe des unités lexicales très variées (*cf.* chapitre 3, p. 67) : des adverbes (« aujourd'hui », « hier »), des syntagmes nominaux (« les années trente », « tous les ans »), des syntagmes prépositionnels (« depuis plusieurs jours », « depuis longtemps », « en 1960 », « des années trente aux années cinquante ») ;
- il fait appel à des traitements de niveaux différents : les prétraitements classiques (segmentation en mots, étiquetage morpho-syntaxique), des lexiques spécifiques, une grammaire ProLog dédiée ;
- nous effectuons un calcul entre le moment de référence temporel de l'expression, la date de publication et la date de lecture (l'année en cours).

L'analyseur temporel assigne aux expressions temporelles repérées une annotation sémantique spécifique. Celle-ci est composée d'une structure de traits qui distingue deux types d'informations : la *nature* de l'expression et sa *situation temporelle* (*sitTps*)[15].

Les valeurs sémantiques concernant la nature et la situation temporelle sont les suivantes :

$$\text{expression temporelle} \left[\begin{array}{ll} nature : & \left[\begin{array}{l} anaphorique \\ deictique \\ duree \\ iteration \\ ponctuel \\ inachevee \end{array} \right] \\ sitTps : & \left[\begin{array}{l} anteriorite++ \\ anteriorite \\ coincidence \\ posteriorite \\ indetermine \end{array} \right] \end{array} \right]$$

L'information sur la **nature de l'expression** est principalement extraite à partir des lexiques (lexique d'adverbes et des prépositions de temps). Dans certains cas, la syntaxe même de l'expression détermine cette caractérisation. Ainsi, la nature *itérative* d'une expression comme « tous les ans » est repérée grâce à une règle comme :

[15]Nous nous sommes inspirée de l'opposition mise en place par Charaudeau (1992) entre « extension temporelle » et « situation temporelle » : alors que dans le premier cas « le processus est considéré d'un point de vue interne, dans sa nature sémantique » (pour l'auteur ponctualité et durée essentiellement), dans le second, « le processus est considéré du point de vue de la position qu'il occupe par rapport à une référence qui correspond à l'instance de parole du sujet ».

```
sp(nature:iteration..sitTps:indetermine) -->
ls_token(_,lemma:tous,word), det(_), ls_token(_,_,nomTemps).
```

Le résultat est de la forme suivante :

(5.1) tous les ans $\begin{bmatrix} nature : iteration \\ sitTps : indetermine \end{bmatrix}$

Dans la plupart des cas, ce sont les informations sémantiques de la préposition (dans les lexiques) qui *remontent* et propagent ces informations sur l'expression temporelle visée. Ainsi, dans les exemples suivants, la valeur *duree* qui est finalement associée à l'expression entière « pendant dix ans » est initialement associée à « pendant » dans le lexique de prépositions de temps. C'est la même chose pour la valeur *ponctuel* qui est liée à « en » ou encore la valeur *inachevee* qui est associée à « depuis ».

(5.2) pendant dix ans $\begin{bmatrix} nature : duree \\ sitTps : indetermine \end{bmatrix}$

(5.3) en 1930 $\begin{bmatrix} nature : ponctuel \\ sitTps : anteriorite++ \end{bmatrix}$

(5.4) depuis 1960 $\begin{bmatrix} nature : inachevee \\ sitTps : anteriorite \end{bmatrix}$

Concernant ce dernier exemple, la figure 5.6 expose la structure de traits associée à la préposition « depuis ». On y lit deux grand types d'informations :

- la valeur de *sitTps* à *inachevee* : elle indique qu'on est dans le cas d'une préposition qui ouvre un intervalle temporel mais que ne le referme pas explicitement ;
- trois informations de type syntaxique : elles informent sur le comportement syntaxique de la préposition. Ici, la préposition se comporte trois façon différentes : *snChiffre* indique qu'elle peut être immédiatement suivie d'un SN de durée (« depuis dix ans »), *snDate* indique qu'elle peut être immédiatement suivie d'une date (« depuis 1990 »),*snAn* indique qu'elle peut être immédiatement suivie d'un SN temporel (« depuis les dernières années »).

L'information sur la **situation temporelle** de l'expression peut également provenir de deux sources différentes :

- remonter du lexique : par exemple, un adverbe déictique comme « aujourd'hui » porte dans son sémantisme le trait *coïncidence* ;
- être le résultat d'un calcul sur la forme de surface lorsqu'il s'agit d'une date chiffrée.

Dans tous les cas, nous distinguons quatre grands intervalles calculés par rapport au moment d'énonciation, *i.e.* du moment de rédaction du texte :

- *antériorité++* : les dates antérieures à 1949
- *antériorité* : de 1950 à 1989
- *coïncidence* : de 1990 à 2008
- *postériorité* : après 2009
- *indéterminé* : les expressions qu'on ne peut pas calculer.

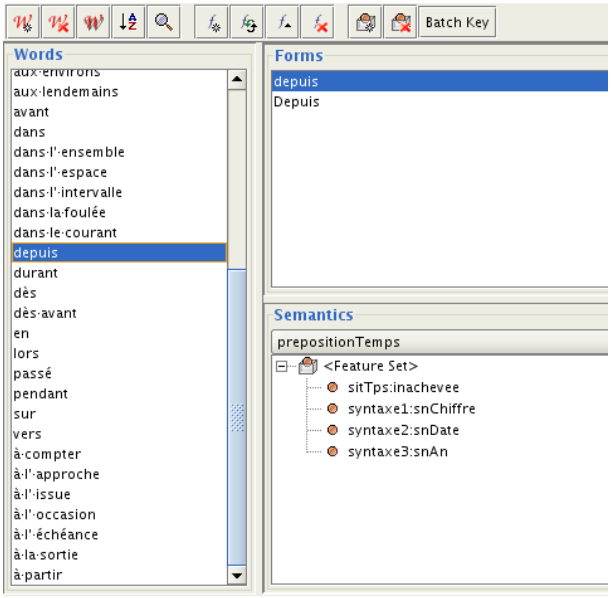


FIG. 5.6 - *Extrait du lexique des prépositions de temps*

Ces intervalles ont été définis suite à la lecture des corpus et en fonction des attentes de mise à jour des rédacteurs. Ils rendent compte d'un système temporel simplifié de la réalité et adapté à notre problématique (*cf.* section 3.2, p. 67).

(5.5) Dès maintenant $\begin{bmatrix} nature : inachevee \\ sitTps : coincidence \end{bmatrix}$

(5.6) En 1960 $\begin{bmatrix} nature : ponctuel \\ sitTps : anteriorite \end{bmatrix}$

5.3.3 Évaluation de l'analyseur des expressions temporelles (temps verbaux et adverbiaux)

Évaluation quantitative

Le nombre d'expressions temporelles repérées est indiqué dans le tableau 5.1 suivant :

	corpus Larousse	corpus Atlas	corpus entier
temps verbaux	4 789	10 979	15 768
adverbiaux temporels	1 651	2 808	4459

TAB. 5.1 - *Nombre des indices de temps*

Le tableau 5.2 rend compte des résultats en termes de précision et de rappel.

	Précision	Rappel
temps verbaux	0.97	0.98
adverbiaux temporels	0.92	0.98

TAB. 5.2 - *Performance de l'analyseur de temps (Précision/Rappel)*

Évaluation qualitative

Concernant les temps verbaux, les résultats sont relativement corrects[16]. Nous avons relevé quelques **bruits** qui sont des cas classiques d'ambiguïtés catégorielles :

(5.7) le greffier [assiste] le juge d'instruction dans ses démarches, [prend] [note] des débats et du déroulement des audiences,

Dans cet exemple « note » est étiqueté comme un verbe au présent.

(5.8) la région du sud-[est] [est] semi-désertique

Ici, les deux formes de surface, identiques, sont également considérées comme des verbes (ce qui est faux pour la première occurrence et vrai pour la seconde).

Nous avons également noté des **silences**. Ces cas sont essentiellement liés à la syntaxe même de la phrase et au fait qu'un mot peut appartenir à plusieurs classes morpho-syntaxiques différentes. Dans l'exemple suivant, le verbe « veille » est considéré comme un nom et n'est pas annoté comme un verbe.

(5.9) il [préside] à [...] et *veille* à

Concernant les adverbiaux temporels, la complexité des traitements ainsi que le caractère imbriqué et interdépendant des analyses nous a amenée à faire certains choix favorisant soit le rappel, soit la précision.

Par exemple, nous avons décidé de ne pas repérer la préposition « Depuis » lorsqu'elle n'est pas suivie d'une virgule (mais elle est repérée lorsqu'elle est immédiatement suivie d'une virgule et lorsqu'elle est suivie d'une date ou d'un SN de temps) :

(5.10) *Depuis* ont été mis sur orbite

Si nous avions décidé de prendre en compte cette unité, l'analyseur aurait repéré des expressions que nous ne voulions pas comme :

(5.11) Une des questions les plus importantes qui se posent au syndicalisme *depuis ses débuts*.

Il y a également un certain nombre de **bruits**. Les exemples suivants rendent compte de cas où l'expression a été repérée alors qu'il ne s'agit pas d'une expression temporelle.

[16] Ils sont comparables aux évaluations connues du TreeTagger (`http://www.ilc.cnr.it/EAGLES/TT-rep/node27.html`)

(5.12) Le capital de la sicav est divisé [en un] certain nombre d'actions.

(5.13) commandées à distance [depuis un] poste d'aiguillage

Lorsque nous avons cherché à éviter ces mauvais repérages, nos modifications entraînaient un taux de silence d'autres adverbiaux temporels élevé. Nous avons donc décidé de conserver ces annotations erronées car cette solution nous semble moins mauvaise que celle qui consiste à ne pas repérer d'autres expressions essentielles. Ce compromis bruit/silence se justifie par rapport à notre objectif qui vise une utilisation humaine de l'application. C'est donc le rédacteur qui jugera en dernier ressort de la validité de l'obsolescence d'un segment. Et il vaut mieux qu'il y ait trop de repérages même s'ils sont erronés que de louper des segments d'obsolescence.

L'exemple suivant, sur la base du mot « mars », rend compte de la question de l'homonymie des mots en français évoquée à la section 5.2.4 (p. 135) : ce mot réfère soit au mois du calendrier, soit au nom d'une planète.

(5.14) L'exploration de [Mars] se poursuit

Enfin, prendre en compte des dates seules entraîne également un nombre de bruit élevé (*a priori*, tous les chiffres auraient été repérés[17]). Ainsi, l'élément suivant n'est volontairement pas repéré.

(5.15) (Philadelphie, *1774*).

5.4 Le traitement de l'aspect

L'analyseur des expressions aspectuelles repère les locutions qui renseignent le lecteur sur le déroulement du procès (« se mettre (enfin) à danser », « des recherches sont en cours », etc.). Comme le montre le schéma 5.7 (p. 143), la grammaire des expressions aspectuelles se base essentiellement sur l'étiquetage morpho-syntaxique (notamment les informations liées au lemme).

Par ailleurs, en français, l'aspect est exprimé par des éléments linguistiques variés comme les verbes ou les adverbes (*cf.* section 3.3 (p.75). Ainsi, par exemple, c'est l'analyseur temporel qui rend compte de la valeur itérative d'un événement.

L'analyseur décrit ici n'a pour ambition que le repérage de périphrases verbales spécifiques.

5.4.1 Le repérage des périphrases verbales

Chacune des expressions repérée est enrichie d'informations qui rendent compte de l'état d'accomplissement du procès : l'accomplissement du procès est évalué à son début, considéré comme terminé ou en cours ou bien il est la négation d'un

[17]Nous aurions pu considérer une date comme une suite de quatre chiffres mais cela entrainait la non prise en considération d'expressions comme « les années 30 » où il s'agit bien d'une date (suite de deux chiffres). Pour améliorer ce système d'annotation du temps, il faudrait des règles contextuelles plus précises.)

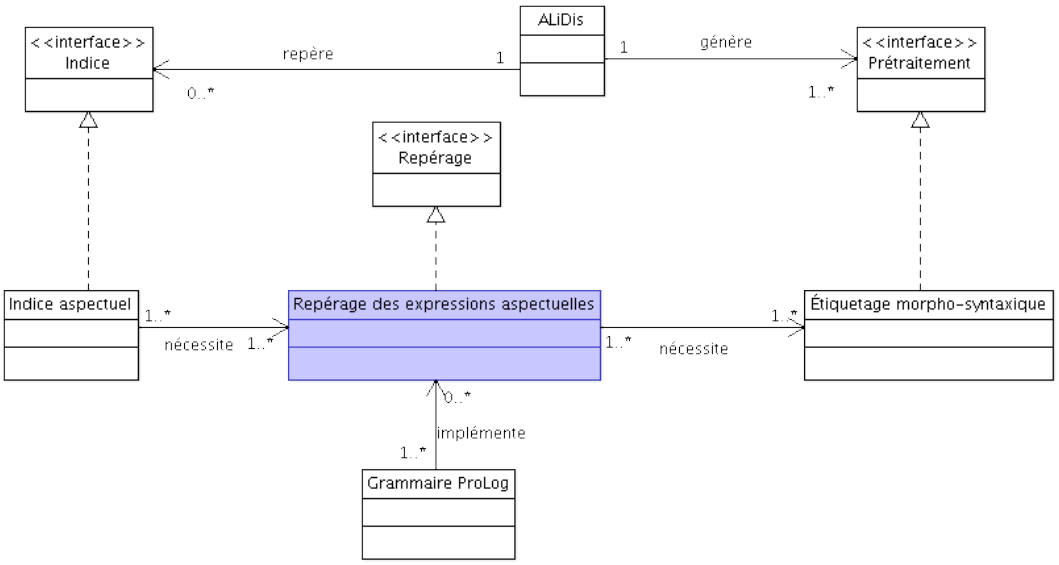


Schéma 5.7 - *Le traitement des expressions à valeur aspectuelle*

procès achevé (« ne pas cesser de »). Les valeurs sémantiques possibles sont au nombre de quatre :

$$\text{expression aspectuelle} \left[\begin{array}{l} \textit{accomplissement} : \left[\begin{array}{l} \textit{debut} \\ \textit{fin} \\ \textit{deroulement} \\ \textit{continuite} \end{array} \right] \end{array} \right]$$

Les exemples suivants illustrent les types d'expressions repérées et annotées par l'analyseur :

(5.16) Les mentalités [...] qui [vont naître] $\left[\textit{accomplissement} : \textit{debut} \right]$ au XVIe s. [...].

(5.17) la société [...] [serait en train de] $\left[\textit{accomplissement} : \textit{deroulement} \right]$ créer une sorte de « gouvernement mondial »

(5.18) Les 3e et 4e tranches [...] [sont en cours de] $\left[\textit{accomplissement} : \textit{deroulement} \right]$ réalisation.

5.4.2 Évaluation de l'analyseur des expressions aspectuelles

Évaluation quantitative

Comme le montre le tableau 5.3, nous repérons un nombre relativement réduit d'expressions de l'aspect.

	corpus Larousse	corpus Atlas	corpus entier
périphrases verbales	30	55	85

TAB. 5.3 - *Nombre d'indices de l'aspect*

Le tableau 5.4 montre les performances de l'analyseur aspectuel en termes de précision/rappel.

	Précision	Rappel
périphrases verbales	0.99	0.43

TAB. 5.4 - *Performance de l'analyseur aspectuel*

Évaluation qualitative

Nous avons focalisé les repérages sur les expressions exprimant un procès en cours ou qui vient de débuter. Pour ces expressions, nous avons relevé quasiment aucun bruit (tout ce qui est prévu dans la grammaire est correctement repéré et annoté).

Le taux de silence est relativement élevé et de nombreuses périphrases verbales marquant l'aspect mériteraient d'être prises en compte. Par exemple, nous ne relevons pas les expressions suivantes qui marquent pourtant le caractère inaccompli d'un procès[18] :

(5.19) il *continue à* se développer

(5.20) la colonisation *se poursuit*

Il est prévu de modifier cet analyseur afin qu'il couvre un plus grand nombre d'expressions aspectuelles.

5.5 Le traitement des entités nommées

L'analyseur d'entités nommées repère et annote les expressions du type de celles présentées dans la section 3.5 (p. 81). Pour rappel, il s'agit principalement de délimiter et annoter les expressions de lieux, les noms propres de personnes, d'organisations (sigles) ou encore les mesures.

Le schéma 5.8 (p. 145) explicite les différents modules requis pour l'analyseur d'entités nommées.

[18] Par manque de connaissances linguistiques et de recul sur les valeurs aspectuelles.

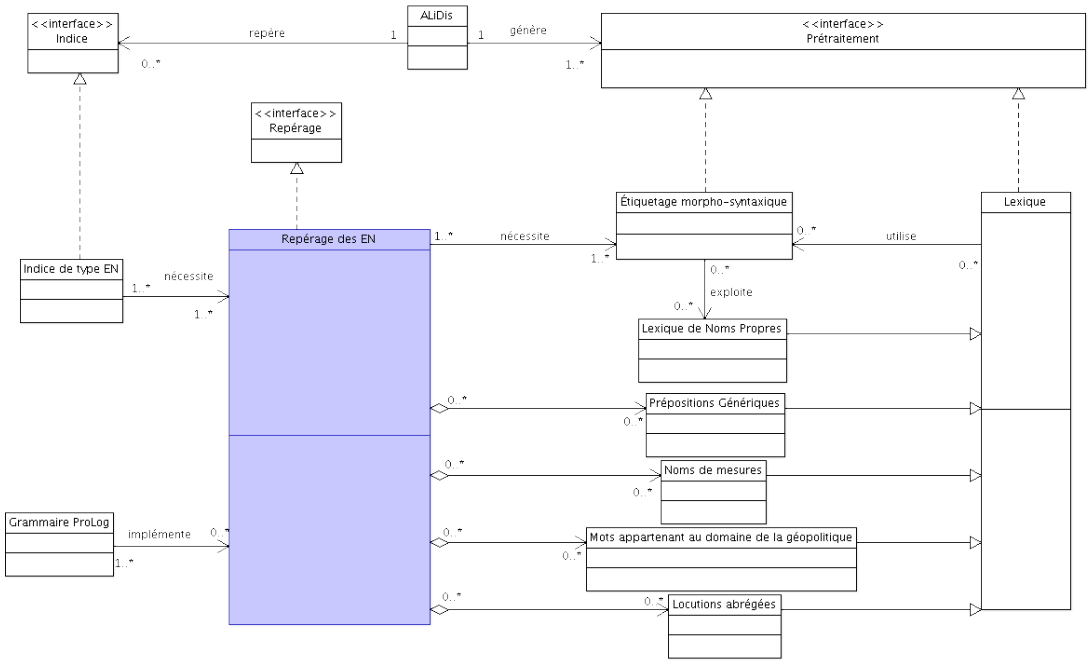


Schéma 5.8 - *Le traitement des entités nommées*

L'analyseur se base sur un certain nombre d'analyses effectuées dans les prétraitements (*cf.* p. 130) : tout ce qui relève des noms propres est désormais disponible à travers le lemme de l'étiquetage morpho-syntaxique (prénoms, noms propres de personnes, sigles, noms propres de pays).

Un lexique de *mots déclencheurs* (*trigger words*) a été constitué. Il a pour objectif le repérage de particules (personne, docteur, société, organisations) et de noms communs de type classifiant (rue, avenue).

Enfin, un lexique de noms de mesure et leurs abréviations ainsi qu'un lexique de termes spécifiques à la géopolitique ont été créés.

Tous les lexiques sont exploités dans une grammaire ProLog. Elle annote les

expressions repérées des valeurs sémantiques suivantes :

$$\text{entité nommée} \left[\begin{array}{l} classe : \left[\begin{array}{l} lieu \\ personne \\ sigle \\ mesure \\ mail/web/marque \\ geopolitique \\ superlatif \end{array} \right] \end{array} \right]$$

Nous décrivons maintenant dans le détail chacune de ces classes définies comme des entités nommées.

5.5.1 L'expression du lieu

Une expression locative peut prendre la forme d'une adresse complète (noms de rue, apposition avec des noms propres, etc.), peut référer à un pays, une ville, une rivière. Syntaxiquement, elle peut être ou bien un syntagme prépositionnel ou bien un syntagme nominal. Cette information de type syntaxique n'est pas nécessaire pour notre objectif général. En revanche, il est intéressant de mettre en évidence les informations sémantiques relatives au type d'expression locative. C'est sur la base ressources créées à partir des données du dictionnaire Larousse que ce type d'information est disponible. Ainsi, une entité nommée de lieu prend l'une des valeurs suivantes :

$$\text{entité nommée de lieu} \left[\begin{array}{l} sousClasse : \left[\begin{array}{l} adresse \\ pays \\ ville \\ riviere \\ pointCardinal \\ indetermine \end{array} \right] \end{array} \right]$$

Voici deux exemples extraits du corpus :

(5.21) entre [São Tomé] $\left[\begin{array}{l} classe : lieu \\ sousClasse : pays \end{array} \right]$ et le [Gabon] $\left[\begin{array}{l} classe : lieu \\ sousClasse : pays \end{array} \right]$

(5.22) Les expéditions d'Arzila [à Tanger] $\left[\begin{array}{l} classe : lieu \\ sousClasse : ville \end{array} \right]$

La valeur *indéterminé* concerne les cas où le mot est étiqueté comme un nom propre par le TreeTagger (et qui n'est pas dans nos lexiques) ainsi que les cas où l'expression est vague.

dans certaines régions $\begin{bmatrix} classe : lieu \\ sousClasse : indetermine \end{bmatrix}$

5.5.2 Les noms de personnes

Les noms de personnes sont repérés à l'aide de lexiques :
- un lexique de noms de personnes
- un lexique de prénoms
- un lexique de particules distinctives (« Mr », « Monseigneur »)

(5.23) La majeure partie du butin ramené par [Drake] $\left[classe : personne \right]$

5.5.3 Les sigles

C'est dans la phase de prétraitement que l'étiquette *sigle* est attribuée aux tokens. L'analyseur d'entités nommées se base alors sur le lemme des tokens qui prend la valeur « @sigle@ »

(5.24) Le trafic de la [VOC] $\left[classe : sigle \right]$ est centré sur le commerce du poivre et d'autres épices.

Nous attribuons également l'étiquette *sigle* pour les noms d'organisation, de société, etc. même s'ils sont écrits dans leur forme complète. Ce choix est lié au fait que notre ressource associe un sigle avec sa forme complète. Ainsi, dans l'exemple suivant, « OCDE » et « Organisation de coopération et de développement économiques » sont associés dans notre lexique de sigles.

(5.25) En 1980, l'[Organisation de coopération et de développement économiques] $\left[classe : sigle \right]$

Nous ne disposons pas en revanche d'information permettant de distinguer finement les sous-classes de sigles possibles (pays, organisation, personne, région, etc.). C'est une évolution qu'il serait intéressant d'apporter à l'analyseur.

5.5.4 Les mesures

Les règles de repérage des mesures utilisent un lexique de noms de mesures (« km », « cm2 », etc.) pour lesquelles on indique si la mesure est plutôt évolutive (les monnaies par exemple) ou statique/fixe (comme les densités ou superficies). Ce choix est le résultat d'une intuition sur le fait qu'il est plus probable qu'une mesure évolue lorsqu'il s'agit de monnaie (*sousClasse :évolutif*) ou de nombre d'habitants (*sousClasse :géopolitique*) que lorsqu'il s'agit de superficie d'un pays

(*sousClasse :fixe*).

$$\boxed{\text{entité nommée de mesure} \left[sousClasse : \left[\begin{array}{l} estimation \\ geopolitique \\ indetermine \\ evolutif \\ fixe \end{array} \right] \right]}$$

(5.26) la montagne Pico de São Tomé ([2 024 m] $\left[\begin{array}{l} classe : mesure \\ sousClasse : fixe \end{array} \right]$).

(5.27) sont [estimées à 60 %] $\left[\begin{array}{l} classe : mesure \\ sousClasse : estimation \end{array} \right]$ de leur budget

(5.28) les litiges supérieurs à [7 600 euros] $\left[\begin{array}{l} classe : mesure \\ sousClasse : evolutif \end{array} \right]$

5.5.5 Les superlatifs

À la section 3.5.4 (p. 83), nous avons évoqué le cas des superlatifs. Nous avons notamment expliqué que, dans le corpus [ENCYCLO] et plus spécifiquement dans le sous-corpus [ATLAS], les superlatifs ont tendance à annoncer la présence d'une entité nommée dans la suite du discours.

Ainsi, une expression comme « le plus haut sommet du monde » annonce l'introduction d'un nom propre de montagne. Nous distinguons deux types de superlatifs : ceux qui indiquent une valeur élevée (« le plus ») et ceux qui indiquent une valeur faible (« le moins »).

$$\boxed{\text{entité nommée de type superlatif} \left[sousClasse : \left[\begin{array}{l} plus \\ moins \end{array} \right] \right]}$$

(5.29) dans [les zones les plus peuplées] $\left[\begin{array}{l} classe : superlatif \\ sousClasse : plus \end{array} \right]$ comme le Bas-Congo

5.5.6 Le domaine de la géopolitique

Enfin, une dernière classe dont nous avons déjà parlé dans la section 3.5.5 (p. 84) concerne les expressions de type « géopolitique » : il s'agit d'éléments comme « densité de population », « taux de mortalité », etc. qui ne sont pas à proprement parler des entités nommées mais qui, comme les superlatifs, les annoncent ou les suivent.

(5.30) Près de 30 % de la [population] $\left[\, classe : geopolitique \,\right]$ est urbaine.

(5.31) Le [taux de chômage] $\left[\, classe : geopolitique \,\right]$ s'est effectivement effondré.

5.5.7 Remarques

L'analyseur repère également les mails, les adresses web et les noms de marque mais nous n'en avons pas trouvé d'occurrences dans le corpus[19].

Une dernière remarque concerne le traitement particulier que nous réservons aux entités nommées dans les titres : nous avons créé deux modules distincts (un pour les paragraphes, l'autre pour les titres) car nous avons constaté que ces expressions ne se réalisent pas forcément exactement de la même manière au sein de ces deux unités. La différence n'est pas fondamentale : la version consacrée aux titres est en fait une version simplifiée de celle consacrée aux paragraphes. Les titres sont en effet des éléments moins complexes que les paragraphes, moins variables dans leur construction. Par exemple, nous n'y trouvons pas de mesures.

5.5.8 Évaluation de l'analyseur des entités nommées

Évaluation quantitative

Le tableau 5.5 indique le nombre d'entités nommées repérées.

	corpus Larousse	corpus Atlas	corpus entier
entités nommées	4 064	8 242	12 306

TAB. 5.5 - *Nombre d'entités nommées dans le corpus* [ENCYCLO]

Le tableau 5.6 montre la proportion des différents types d'entités nommées traités.

entités nommées de lieu	37.9 %
entités nommées de personne	12.7 %
entités nommées de type sigle	8.8 %
entités nommées de type mesure	15.8 %
entités nommées de type géopolitique	20.6 %
entités nommées de type superlatif	4.3 %

TAB. 5.6 - *Proportion des types d'entités nommées dans le corpus* [ENCYCLO]

Le tableau 5.7 montre les performances de l'analyseur d'entités nommées.

[19] Nous avons décidé d'écrire des règles pour les repérer car nous avions remarqué que certaines fiches Atlas faisaient référence à des sites Web : un employé de la société INITIALES était d'ailleurs chargé de vérifier les liens Internet et de les mettre à jour. Les fiches sur lesquelles nous travaillons n'en contiennent pas.

	Précision	Rappel
entités nommées	0.99	0.83

TAB. 5.7 - *Performance de l'analyseur d'entités nommées*

Évaluation qualitative

Les noms propres présents dans les lexiques sont naturellement bien repérés s'il sont également présents dans le lexiques des noms propres. Le taux de silence est donc lui aussi lié au lexique et à l'absence du terme dans ce lexique.

Il y a un certain nombre d'erreurs d'annotations liées à une mauvaise attribution d'une valeur sémantique dans les lexiques ou liées à des erreurs d'étiquetage.

Dans l'exemple 5.32, « Ans » est annoté comme une ville car le mot est présent dans notre lexique de noms propres et étiqueté comme tel. Mais dans ce cas, il ne s'agit bien entendu pas d'une ville mais d'une période de l'histoire. Dans l'exemple 5.33, « Donne » est annoté comme un nom propre de personne.

(5.32) la guerre de Sept [Ans]

(5.33) La politique de New Deal (« Nouvelle [Donne] »)

Certaines expressions ne sont volontairement pas repérées parce qu'elles ont tendance à bruiter considérablement les résultats[20]. C'est le cas notamment pour les éléments chiffrés qui n'ont pas d'objet clairement identifiable (nous ne traitons pas les anaphores) :

(5.34) Ils sont *181* en France

(5.35) Au nombre de 7000, ils sont [...]

Par ailleurs, dans de nombreux cas, c'est un repérage partiel de l'expression qui est effectué. Dans l'exemple suivant, seul le nom propre « Paris » est repéré et annoté.

(5.36) *À la Bourse de* [Paris],

La situation est identique pour les coordinations qui ne sont pas non plus repérées :

(5.37) huertas du [Levant] et de l'[Andalousie]

Les répercussions de ces annotations partielles ne sont pas centrales pour notre tâche puisque les informations sémantiques pertinentes sont malgré tout représentées (en double). C'est plus pour rendre les traitements plus propres qu'il serait souhaitable d'utiliser un outil d'annotation syntaxique comme Syntex[21]. Un tel

[20] Ce problème est proche de celui déjà présenté pour l'analyseur temporel, *cf.* section 5.3 (on ne repère pas les dates seules pour les mêmes raisons).

[21] `http://w3.erss.univ-tlse2.fr/textes/pagespersos/bourigault/syntex.html`

outil permettrait de considérer systématiquement les syntagmes dans leur ensemble. mais nous restons prudents quant à l'amélioration globale du système par un outil d'annotation syntaxique. Il serait intéressant de tester cette possibilité qui pourrait d'ailleurs servir à d'autres repérages et annotations comme la résolution d'anaphores par exemple.

5.6 Le traitement de la modalité

Le schéma 5.9 explicite que l'analyseur de la modalité se base essentiellement sur trois prétraitements : il exploite les formes tokenisées (notamment pour l'utilisation de la ponctuation), il utilise des informations morpho-syntaxiques ainsi que des lexiques : lexiques de sources d'information (« rapport », « recherche »), d'adverbes médiatifs et thématiques (« Selon », « Concernant »), d'adverbes affectifs (« malheureusement », « sans doute »), de noms, adjectifs et verbes liés au point de vue du rédacteur (« hypothèse », « improbable », « penser ») et un lexique de mots argumentatifs (« donc », « par exemple »).

De plus, cet analyseur exploite certaines annotations produites par l'analyseur d'entités nommées dont nous avons déjà parlé : il s'agit des sigles notamment.

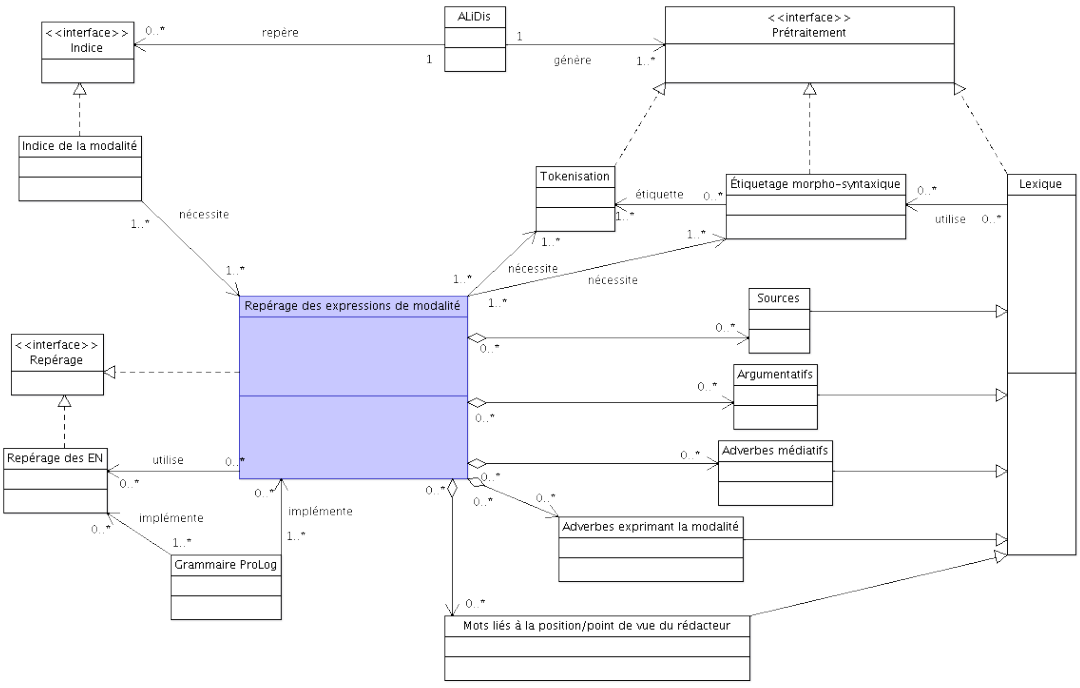


Schéma 5.9 - *Le traitement des expressions de la modalité*

L'analyseur est basé sur les observations décrites dans le chapitre 3.4 (p. 77) dans lequel sont opposés deux grands types de modalités : la modalité d'énoncé et la modalité d'énonciation.

5.6.1 La modalité d'énoncé

Un premier programme a pour objectif de déterminer le type de l'énoncé : il s'agit, pour chaque phrase du corpus de déterminer, sur la base de la ponctuation,

si elle est assertive (« . », « ... »), exclamative (« ! ») ou interrogative (« ? »). Et s'il n'y a pas de ponctuation, alors aucune annotation n'est effectuée.

Le tableau 5.8 indique le nombre de phrases assertives, exclamatives et interrogatives dans le corpus [ENCYCLO]. On constate que les phrases exclamatives et interrogatives sont très peu nombreuses.

	Corpus [ENCYCLO]
Phrases assertives	9618
Phrases interrogatives	23
Phrases exclamatives	3

TAB. 5.8 - *Nombre de phrases assertives, exclamatives et interrogatives dans le corpus* [ENCYCLO]

5.6.2 La modalité d'énonciation

Le second programme est plus complexe notamment parce qu'il repère et annote des expressions assez diverses. La structure de traits associée à chacune d'elles peut prendre une des valeurs suivantes :

$$\text{expression de la modalité} \left[\textit{type} : \left[\begin{array}{l} \textit{jugement} \\ \textit{recence} \\ \textit{prevision} \\ \textit{importance} \\ \textit{distance} \\ \textit{jugePerso} \\ \textit{source} \\ \textit{definition} \\ \textit{restriction} \\ \textit{enonciatif} \\ \textit{thematique} \end{array} \right] \right]$$

Cette classification a été mise en place de manière intuitive (et parfois arbitraire) en relation avec la tâche visée et sur la base du corpus annoté manuellement. Les classes ont ensuite été enrichies sur la base de dictionnaires de synonymes et d'informations fournies par des grammaires du français (Riegel *et al.* (1994) principalement). La constitution de cette ressource mériterait d'être affinée selon une catégorisation plus précise et mieux adaptée à notre tâche. C'est également parce que nous n'avons pas trouvé de ressource numérique disponible que nous l'avons créée.

Le premier type (***jugement***) concerne les adverbes de modalité susceptibles de traduire une émotion (« heureusement »), un doute (« peut-être ») de la part du locuteur, etc.

(5.38) La prolifération des États signataires de la Charte des Nations unies renforce [peut-être] $\left[\, type : jugement \,\right]$ l'[impression] $\left[\, type : jugement \,\right]$ d'homogénéité juridique de la communauté internationale.

Nous incluons dans la modalité des expressions qui tendent vers une **interprétation temporelle** et qui ne s'interprètent que si l'on connaît la date de publication du texte. Ces repérages ne sont pas développés dans le traitement du temps mais dans celui de la modalité car nous considérons que la valeur temporelle est adjointe à un sémantisme principal non-temporel. Elle n'est pas ici centrale dans la propos de l'auteur. Elle est véhiculée par un adjectif temporel qui qualifie un nom commun principal. Les trois exemples suivants illustrent ces cas :

(5.39) soit [les territoires actuels] $\left[\, type : recence \,\right]$ de la Région île-de-France et le sud de la Région Picardie

(5.40) le [dernier Mondial] $\left[\, type : recence \,\right]$ s'est tenu à

(5.41) les [recherches à venir] $\left[\, type : prevision \,\right]$

Le rédacteur peut également insister sur l'**importance d'un fait** à un moment donné :

(5.42) Il s'agit d'[un véritable enjeu] $\left[\, type : importance \,\right]$

Les expressions permettant au rédacteur de se distancier des propos qu'il relate sont également repérées dans nos corpus.

Certaines expressions rendent compte de la capacité de l'auteur à présenter **différents points de vue** pour un même fait.

(5.43) on estime $\left[\, type : distance \,\right]$

(5.44) il pense $\left[\, type : distance \,\right]$

Les outils linguistiques de distanciation (les **marqueurs évidentiels**) peuvent également introduire des propos proclamés par des instances connues, des personnalités reconnues :

(5.45) Selon le rapport de l'INSEE $\left[\, type : source \,\right]$

L'analyseur prévoit de repérer les cas où le rédacteur utilise la première personne du singulier avec un **verbe de jugement** (« je pense »). Nous n'avons pas trouvé d'occurrence de ce type d'expressions qui dans ce cas auraient été annotées avec la structure de traits $\left[\, type : jugePerso \,\right]$.

Certaines formes introduisant une **définition** sont également repérées. Nous avons remarqué que les définitions, qui sont des objets souvent explicites et plus aisés à repérer automatiquement, n'apparaissent pjamais dans les segments d'obsolescence. C'est pour nous permettre de repérer également ce qui n'est pas obsolescent que nous choisissons de repérer des éléments comme les définitions :

(5.46) On distingue $\left[\, type : definition \,\right]$ deux classes :

Enfin, concernant la capacité des rédacteurs à organiser leur pensée, nous classons dans la modalité les **éléments argumentatifs** comme « d'abord », « puis », « dans un premier temps », etc., des expressions comme « À ce sujet/propos » qui

est annoté $\left[\, type : enonciatif \,\right]$ ou encore « Pour ce qui est de la dette » qui est annoté $\left[\, type : thematique \,\right]$. Ce sont des outils linguistiques permettant à l'auteur d'organiser sa pensée, de structurer ses propos.

Évaluation de l'analyseur des expressions de la modalité

Évaluation quantitative

Le nombre d'expressions de la modalité (d'énonciation) que l'analyseur repère est indiqué dans le tableau 5.9.

	corpus Larousse	corpus Atlas	corpus entier
expressions de point de vue	330	586	916

TAB. 5.9 - *Nombre d'indices de la modalité*

Le tableau 5.10 rend compte de la performance en termes de précision et de rappel de l'analyseur des expressions modales.

	Précision	Rappel
expressions de point de vue	0.73	0.98

TAB. 5.10 - *Proportion et évaluation des indices de la modalité*

Évaluation quantitative

Cet analyseur produit un certain nombre d'expressions bruitées. Ceci est principalement lié à la nature polysémique de certaines formes. Ainsi, l'exemple suivant n'est pas une prédiction sur l'avenir interprétée selon le point de vue temporel du rédacteur.

(5.47) la loi [prévoit]

L'exemple suivant ne rend pas compte d'un jugement humain :

(5.48) les règles écrites sur [la chose jugée]

Pour résoudre ce type de problèmes, il faudrait certainement contraindre la règle pour qu'elle n'accepte que des sujets de type humain.

Nous constatons un problème similaire dans l'exemple suivant où une contrainte sur le nom ou sur la préposition pourrait sans doute améliorer l'annotation :

(5.49) en [dernier recours] / en [dernier ressort]

Enfin, utilisant un système où chaque analyse dépend étroitement des analyses précédentes, nous avons constaté que des erreurs d'étiquetage morpho-syntaxique entraînent des erreurs dans les résultats de nos analyseurs. Dans l'exemple suivant, le mot « concurrence » est un verbe et non un nom comme l'a étiqueté TreeTagger.

(5.50) La vitesse de transmission des données sur l'Internet fait que [ce dernier concurrence] les moyens traditionnels

Ci-après, les majuscules auraient dû contraindre un étiquetage en termes de nom propre.

(5.51) la [Voie Moderne]

Dans les exemples suivants, les expressions n'ont pas été repérées soit parce que l'expression n'est pas prise en considération (exemple 5.52), soit parce qu'il y a eu une erreur de tokenisation ou d'étiquetage morpho-syntaxique (exemple 5.53) :

(5.52) *est menacée* par la déforestation

(5.53) de *nouvelles études* seront menées

5.7 Le traitement de la position des indices dans la phrase

Cet analyseur exploite l'ensemble des repérages (expressions de temps, de modalité, d'aspect et les entités nommées) dont nous venons de parler ainsi que le module de délimitation des phrases (*cf.* section 5.2.3, p. 132). Il a été élaboré à l'aide du composant MRE (Macro Regular Expression) de LINGUASTREAM. Ceci est représenté dans le schéma 5.10.

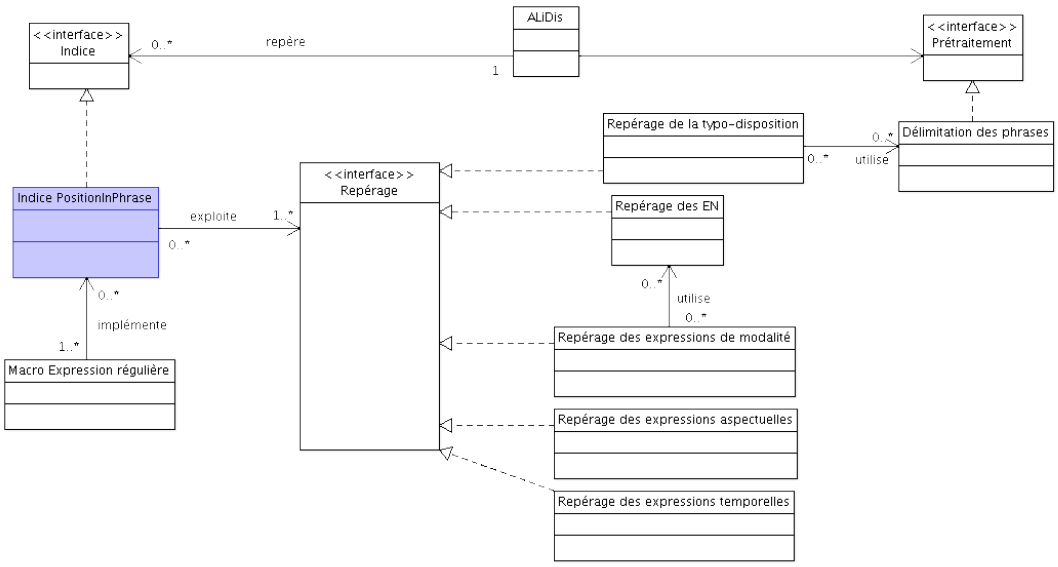


Schéma 5.10 - *Le traitement de la position des indices dans le texte*

Nous considérons un début de phrase comme une initiale stricte, c'est-à-dire qui suit immédiatement la balise XML ouvrante concernant la délimitation de la phrase. La situation est identique pour la position finale. La position *Amorce* est également traitée pour des cas comme :

(5.54) La [population :] le nombre d'habitants ne cesse d'augmenter.

Les résultats de cet analyseur dépend exclusivement de la bonne délimitation des phrases et du repérage correct des indices. Ainsi, l'exemple suivant illustre un cas où la position initiale n'est pas repérée car l'indice temporel est mal repéré. L'expression temporelle n'étant que partiellement repérée, elle n'est donc pas considérée comme étant située à l'initiale (stricte) de la phrase.

(5.55) Dans la première [moitié du XVIIe siècle],

5.8 Exploitation de la structure XML et des méta-données

Cet analyseur (*cf.* schéma 5.11) écrit en macro-expressions régulières (MRE) exploite des informations textuelles de granularité élevée : il s'agit de prendre en compte les informations transmises au-delà de la proposition et jusqu'à l'ensemble du texte voire du corpus. C'est également lui qui gère les titres à travers la prise en compte du balisage XML initial.

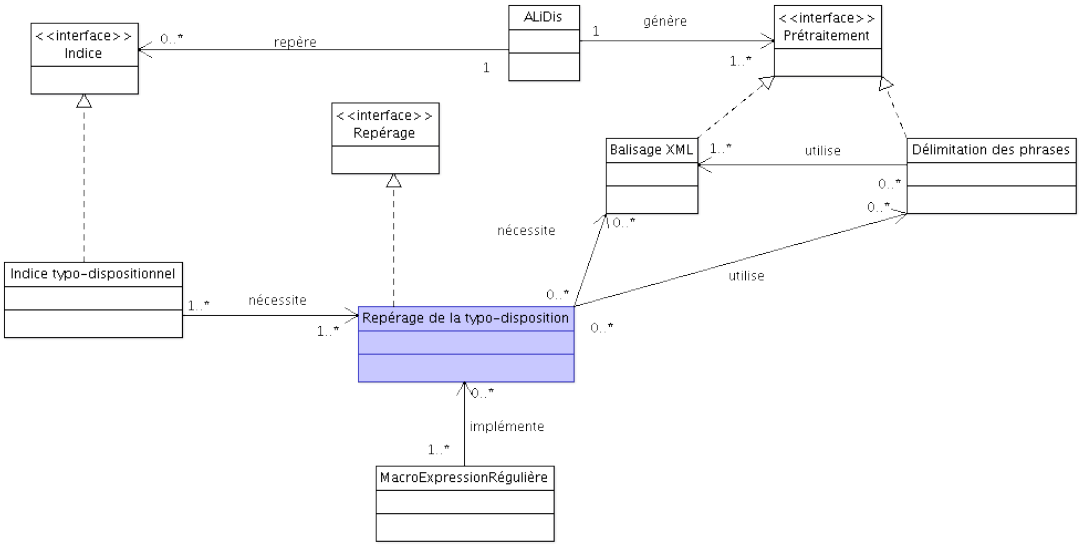


Schéma 5.11 - *Le traitement de la typo-disposition*

À partir de la structure XML des documents et du module de délimitation des phrases, nous exploitons :

- la position des phrases au sein du paragraphe ;
- la position du paragraphe au sein de sa partie : en position d'introduction et en position de conclusion sur les parties de niveau 1 (*debutZone* et *finZone*) et de niveau 2 (*debutDivision* et *finDivision*) ;
- la caractérisation en tant que titre des éléments textuels.

Pour éviter qu'une phrase soit étiquetée à la fois comme étant à l'initiale du paragraphe et en finale (lorsqu'elle est seule dans le paragraphe), nous attribuons à cette phrase le trait sémantique *seuleInParagraphe*.

Enfin, c'est également dans cet analyseur que la *thématique* des textes est prise en compte. En effet, cette information est disponible comme attribut dans le balisage XML.

5.9 Récapitulatif général

Le tableau 5.12 résume l'ensemble des indices que nous considérons ainsi que les structures de traits associées. Il rend compte de la variété des types d'indices pris en compte et de leur variation en termes de granularité dans le document.

Types de marqueurs	**Structure de traits**		**Implémentation**
	Trait 1	**Trait 2**	
Adverbiaux temporels	nature : *ponctuel, inachevee, deictique, duree, iteration*	sitTps : *anteriorite++, anteriorite, coincidence, posteriorite, indetermine*	Lexiques + Grammaire
Temps verbaux	temps : *passeComp, passeAnt, plusQuePft, futAnt, condPasse, present, passeSimple, imparfait, futur, conditionnel*		TreeTagger + Grammaire
Périphrases verbales	accomplissement : *debut, fin, deroulement, continuite*		TreeTagger + Grammaire
Entités nommées	classe : *personne, lieu, sigle, web, mail, marque, geopolitique, mesure*	sousClasse : *riviere, pays, ville, evolutif, fixe,...*	Lexiques + Grammaire
Expressions du point de vue	type : *distance, jugement, recence, prevision, importance, jugePerso, source, thematique, restriction*		Lexique + Grammaire
Argumentatifs	type : *correction, explication, opposition, conséquence, temporelle, exemplification,...*		Lexique
Type de phrase	type : *exclamation, assertion, interrogation*		MRE
Position de l'indice dans la phrase	debut, fin, amorce		MRE
Position de la phrase dans le paragraphe	position : *debutParag, finParag, seuleInParagraphe*		MRE
Position du paragraphe dans le document	position : *debutZone, finZone, debutDivision, finDivision*		MRE

Schéma 5.12 - *Résumé des marqueurs textuels et discursifs repérés automatiquement*

Le tableau 5.11 indique le nombre total d'occurrences des indices ainsi que leur évaluation en termes de précision/rappel. Les résultats sont satisfaisants et nous autorisent à mener un apprentissage automatique sur nos données annotées manuellement (les segments d'obsolescence) et annotées automatiquement (les indices linguistiques).

Pris isolément ces traitements linguistiques peuvent paraître approximatifs (notamment en termes des bruits et silences constatés). Il faut toutefois insister sur le fait qu'il s'agit d'une étude exploratoire sur le phénomène de l'obsolescence : les résultats et observations que nous faisons dans le chapitre 8 (p. 211) permettront de

	corpus Larousse	corpus Atlas	corpus entier
Nombre Total des indices	16 070	35 294	51 364

TAB. 5.11 - *Nombre total des indices*

	Précision	Rappel
Performance multiclasse	0.93	0.85

TAB. 5.12 - *Performance globale de l'outil* ALIDIS

revenir sur le repérage de ces indices, de les affiner, les supprimer ou les modifier afin qu'ils soient les plus pertinents possible pour caractériser les segments d'obsolescence. De plus, ces indices n'ont pas vocation à être utilisés de manière isolée : nous recherchons des combinaisons d'indices qui, elles, présentent une valeur particulière pour l'obsolescence.

Les erreurs d'annotations constatées sont dans la plupart des cas contrôlées et présents à notre esprit pour l'interprétation future des résultats statistiques.

« L'une des qualités appréciables d'un instrument par rapport à un annotateur humain, c'est que son comportement est reproductible. Face aux mêmes données, il reproduira toujours le même résultat. Utiliser au mieux un instrument, c'est donc profiter de cette stabilité pour savoir quels sont ses biais, ses réussites et ses erreurs systématiques. » (Habert, 2005, p39)

Enfin, nous souhaitons conclure sur le caractère entièrement reproductible et réutilisable de l'outil ALIDIS : s'il est aujourd'hui possible de modifier les analyseurs et les ressources actuelles, il est également possible d'envisager d'ajouter d'autres traitements, d'autres analyses de façon relativement simple[22]. Par exemple, nous serions vivement intéressée par le développement d'un analyseur d'expressions anaphoriques (antécédents des pronoms, ellipses, références) ou de la détermination de la structure rhétorique (commentaires, explications, causalités, etc.).

[22]En partant du principe que l'utilisation de la plateforme LINGUASTREAM est maîtrisée.

Chapitre 6

Étape 2 : Outil OCAS (Outil de Création d'Abstraction Sémantique)

L'outil OCAS a pour objectif la transformation de données *brutes* (*i.e.* les sorties XML de LINGUASTREAM) en données *traitables* par des outils statistiques. Pour cela, OCAS est basé sur un modèle conceptuel des données dans lequel les imbrications hiérarchiques et relationnelles des données textuelles (annotations manuelles et automatiques) sont représentées[1].

OCAS fait le lien entre un modèle linguistique (*i.e.* l'ensemble des indices repérés et annotés par ALIDIS) et un modèle statistique *via* un modèle conceptuel de données. Ce modèle conceptuel de données permet de rendre compte de l'imbrication des éléments entre eux (hiérarchie entre les indices) et de leurs associations (relations entre les indices). Cette question du passage des données textuelles (XML) en données numériques est centrale dans toute étude en discours. Il est cependant encore mal résolu. La solution que nous proposons est une réponse partielle à ce problème. S'il est évident que nos choix sont discutables et ont des limites, ce modèle nous permet malgré tout d'avancer dans notre tâche.

Le schéma 6.1 explicite le fonctionnement de l'outil OCAS : il nécessite en entrée un corpus annoté (manuellement ou/et automatiquement) et propose en sortie ce que nous dénommons une *abstraction sémantique des textes*. Cette abstraction peut, selon les besoins, prendre la forme d'une base de données, d'un fichier CSV, RTF, XML, EXCEL, etc.

L'abstraction sémantique cherche à représenter le plus fidèlement possible la réalité des données linguistiques brutes. Elle est cependant établie sous une forme différente : les éléments sont projetés d'un plan a (le texte annoté) à un plan b (des éléments dans une base de données) sans perte d'information. Ce qui n'empêche pas que l'information soit transformée.

[1] Nous remercions Frédéric Gardes pour l'aide apportée dans la création de cet outil, tant au niveau de sa conception que de sa réalisation informatique.

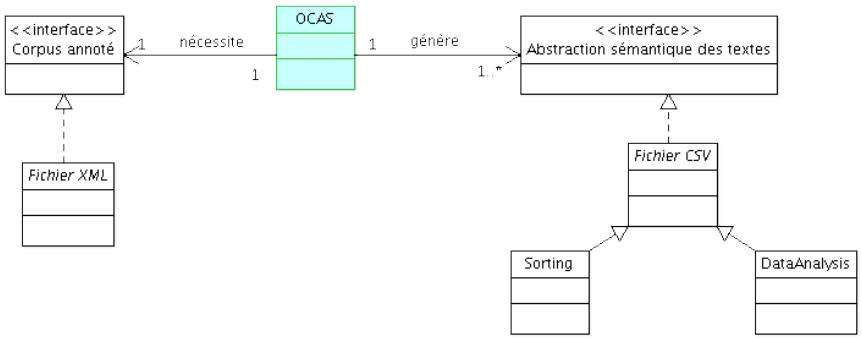


Schéma 6.1 - *L'outil OCAS*

Pour rappel, les données issues de l'outil ALIDIS respectent la classification (linguistique) suivante :

- Les **éléments intra-phrastiques** font référence aux indices internes à la phrase.
- Les **caractéristiques positionnelles** rendent compte à la fois de la position des indices intra-phrastiques à l'intérieur de la phrase et de la présence de ces mêmes indices dans un titre. Cet indice de type positionnel requiert donc nécessairement le repérage préalable d'un autre indice de granularité plus petite. Nous distinguons les titres en les classant comme des **indices hiérarchiques**. En effet, leur comportement sera différent des indices positionnels internes à la phrase, car leur rôle dans le document (position titre, au delà de la phrase et même du paragraphe) est particulier.
- Les **indices de position discursive** rendent compte de la position des unités phrases au sein des paragraphes (première phrase ou dernière phrase du paragraphe) et des unités paragraphes au sein du document (premier paragraphe ou dernier paragraphe de la section, sous-section, etc.).
- Les **indices externes** concernent le type de document ou les informations sur le domaine auquel le texte appartient.

Par ailleurs, les indices linguistiques pris en compte sont de types variés (repérage du temps, de la modalité, de l'aspect, des entités nommées, de la position des indices dans la phrase, dans un titre, dans le paragraphe ou dans le document, etc.).

La question de la variabilité du grain d'analyse est centrale pour la détermination des individus à prendre en compte. Quant aux types d'indices, ils sont directe-

ment liés à la catégorisation et la caractérisation des variables. La complexité de ces analyses à granularité variée nous a posé de nombreux questionnements quant à la manière de les exploiter et nous a amenée à construire un modèle de représentation des données non trivial.

6.1 Le modèle de représentation des données

Le schéma 6.2 présente le modèle construit pour représenter de manière uniforme l'ensemble des données textuelles et linguistiques du corpus.

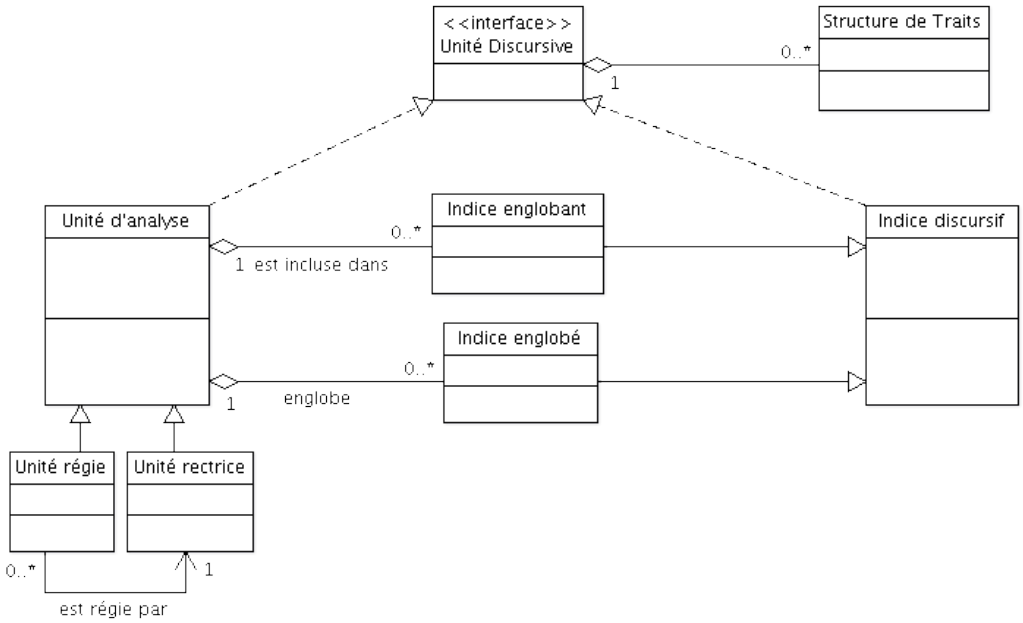


Schéma 6.2 - *Le modèle conceptuel des données*

Dans ce modèle, le premier point important est que tout élément textuel peut être considéré comme une unité discursive, qu'il s'agisse d'une unité d'analyse, *i.e.* l'unité d'analyse de base pour les traitements, ou d'un indice discursif, qu'il soit englobant ou englobé. Une unité discursive est caractérisée par un type (par exemple le type « expression temporelle » ou le type « phrase »).

Le second point important est que toute annotation peut potentiellement être une *unité d'analyse*. N'importe quel indice, qu'il soit intra-phrastique (une expression temporelle), positionnel (un introducteur de cadre) ou à grande granularité (un paragraphe) peut prétendre à devenir une unité d'analyse. Cette souplesse relative à la granularité de l'unité d'analyse est un concept hérité de la philosophie et de la

conception de la plateforme LINGUASTREAM : il est possible, à tout moment, de modifier le grain d'analyse souhaité. Ce besoin de souplesse est intrinsèquement lié au fait que, dans des études en discours, les phénomènes linguistiques étudiés sont parfois hétérogènes (par exemple, l'étude des cadres de discours implique la prise en compte de segments de granularité variable). De plus, notre approche est exploratoire quant à la caractérisation du phénomène de l'obsolescence. On ne peut pas savoir *a priori* quel sera le bon grain d'analyse. L'unité pourrait tout aussi bien être le mot, la phrase ou le paragraphe ou même deux de ces modalités. Nous avons donc besoin d'un modèle qui nous permette de changer facilement de granularité.

Chaque **unité discursive** qu'elle soit unité d'analyse ou indice discursif, est potentiellement rattachée à une structure de traits. Le nombre de traits de la structure est également généré dynamiquement en fonction des données textuelles, et non précisé *a priori*. La structure de traits peut ne comporter aucun trait, ou bien deux ou trois ou dix, le modèle gère cette inconnue.

Concernant l'unité d'analyse, elle est potentiellement reliée :

- **à un ou plusieurs indices discursifs qui l'englobe :** si on considère que la phrase est l'unité d'analyse, alors les indices englobants sont les **indices positionnels** (par exemple le paragraphe dans lequel elle se réalise). Étant donné qu'on ne peut jamais connaître le nombre d'unités englobantes pour une unité d'analyse donnée, l'outil est conçu de telle sorte que les relations soient gérées dynamiquement. Le nombre d'unités discursives englobantes n'est jamais déterminé *a priori* et on peut ajouter des indices discursifs à l'infini (en quantité mais aussi en types différents). Dans notre travail, les **annotations manuelles de l'obsolescence** sont considérées comme des indices discursifs englobants.
- **à un ou plusieurs indices discursifs qui est inclus :** si on considère toujours la phrase comme unité d'analyse, alors les indices englobés correspondent aux **indices intra-phrastiques** (indices temporels, aspectuels, les entités nommées, etc.). Comme pour les unités englobantes, leur génération est dynamique et non fixée à l'avance et on peut également en rajouter autant que nécessaire. À noter que les **indices positionnels intra-phrastiques** font également partie de ce type d'unité[2].
- **à une ou plusieurs autres unités d'analyse qui la régissent :** c'est la solution adoptée pour traiter le cas des relations entre des unités d'analyses différentes. Dans notre cas, il s'agit du traitement des **indices hiérarchiques** (*i.e.* les titres). Les titres sont des unités qui fonctionnent à la fois comme les unités phrases (ils sont susceptibles d'être reliés tant à une unité englobante qu'à une unité englobée) mais qui présentent la particularité de ne pas être obsolescents. Nous avons notamment insisté sur leur potentialité à prédire l'obsolescence.

Ce modèle propose de répondre à trois problématiques centrales :

[2]Nous envisageons une évolution du modèle permettant de considérer les introducteurs de cadre (IC) comme des indices englobant (*cf.* explications p. 167).

1. le fait qu'une unité d'analyse puisse être reliée à un certain nombre d'indices englobants et à un certain nombre d'indices englobés (*i.e.* l'*associativité*).
2. la question du nombre d'indices pour un individu donné (*i.e.* la *scalabilité* ou *extensibilité*, soit 1 individu avec n indices) : ce nombre n étant inconnu à l'avance (par exemple, on ne peut pas savoir combien de verbes contient une phrase), il nous faut un système capable de gérer dynamiquement l'attribution du nombre d'indices pour chacune des unités d'analyse ; cette question se pose également pour la gestion des structures de traits.
3. le fait qu'une unité d'analyse dépende d'une autre unité d'analyse (*i.e.* la *réflexivité*) : dans notre travail, il s'agit notamment de la question des titres et de leur relation avec les phrases. Cette notion rejoint celle d'héritage du contexte de Zerida *et al.* (2006) dont nous avons parlé dans la section 1.3.3 (p. 32).

6.1.1 L'associativité

Nous insistons ici sur la capacité pour chacune des unités d'analyse à être associée à n'importe quel type d'indice discursif (indice englobant et indice englobé) ou une autre unité d'analyse. Ainsi, sont associées entre elles des unités discursives différentes et qui, dans la réalité, sont parfois hiérarchiques ou qui peuvent se chevaucher.

6.1.2 L'extensibilité

Ici se pose la question du nombre d'indices englobants et englobés par unité d'analyse et du nombre de traits sémantiques par unité discursive.

Pour une unité d'analyse donnée, le nombre et le type d'indices englobants et englobés ne peut pas être prévu *a priori*. Il s'agit donc de résoudre la question du nombre puisqu'on ne peut pas définir à l'avance la quantité d'indices discursifs (du même type ou de types différents) susceptibles d'englober ou d'appartenir à une unité d'analyse. C'est ce que représente le schéma 6.3 suivant.

Cette problématique du nombre se retrouve dans le traitement des structures de traits de chacune des unités discursives. Les programmes en amont sont conçus pour qu'il y ait au maximum deux traits pour une structure de traits mais il est tout à fait envisageable d'en augmenter le nombre si cela est nécessaire.

La solution proposée permet de gérer les individus de manière entièrement dynamique sans avoir à déterminer *a priori* un nombre et/ou un type prédéfini d'indice pour un individu et un nombre et/ou un type de traits prédéfini pour un indice. Il s'agit de *laisser parler* le corpus en travaillant uniquement sur les occurrences effectives des expressions linguistiques repérées.

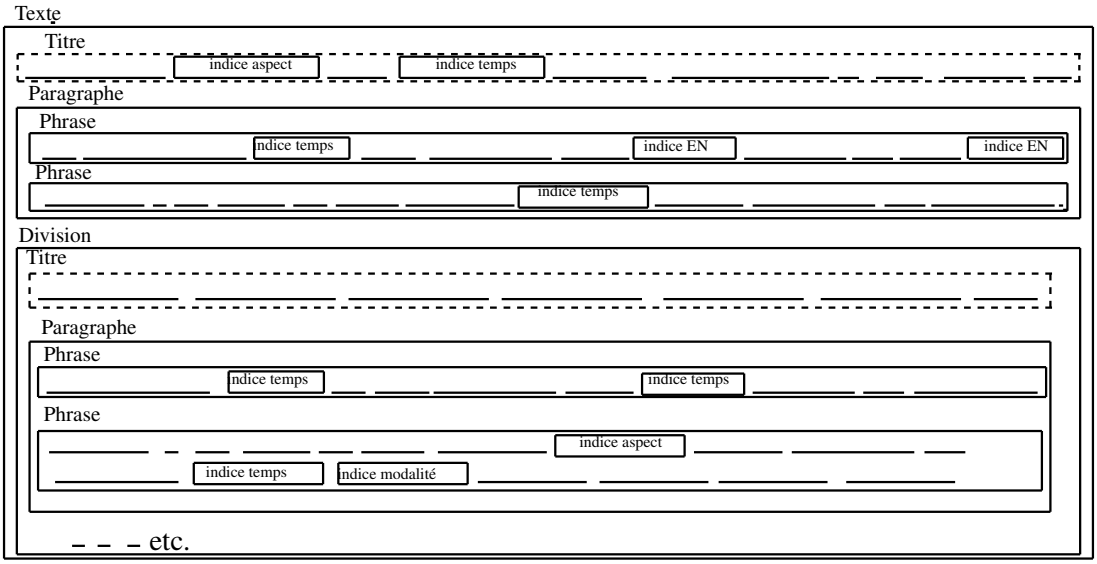


Schéma 6.3 - *L'imbrication des unités d'analyse*

6.1.3 La réflexivité

Dans notre travail, les phrases et les titres sont des unités d'analyse. Une phrase est régie par un titre : il y a une relation particulière entre ces deux unités d'analyse. Nous parlons d'unité d'analyse régie (la phrase) et d'unité d'analyse rectrice (le titre).

Un titre est annoté par les mêmes indices que ceux des phrases (temps, aspect, modalité et entités nommées). Mais, à la différence de la phrase, un titre ne peut pas être obsolescent[3]. Nous le supposons cependant bon prédicteur de la présence de segments obsolescents dans sa partie (*cf.* section 4.3.3, p. 105).

Nous avons donc fait le choix de projeter chacune des informations sémantiques (*i.e.* les structures de traits) des indices contenus dans les titres sur chacune des phrases régies par le titre en question.

Cette solution permet de prendre en compte les informations véhiculées par les titres et de les envisager comme éléments des combinaisons d'indices susceptibles d'aider au repérage des segments d'obsolescence.

[3]Cette condition est définie dans le protocole d'annotation manuelle, p. 46.

6.2 Limites

Perte de hiérarchisation et des relations

Ce modèle ne rend pas compte de la hiérarchie interne ou des relations linéaires qui existent potentiellement entre les indices englobés ou les indices englobants. L'ensemble des indices associés à une unité d'analyse est remis à plat, sans relation entre eux.

Les configurations d'indices au-delà de l'unité phrase

Un autre problème important concerne le fait que les combinaisons d'indices qui vont être apprises concernent des indices liés à un même individu, à savoir la phrase puisqu'il s'agit de l'unité d'analyse sur laquelle nous travaillons. Or parfois, il est intéressant de pouvoir considérer des configurations d'indices sur un ensemble indéterminé de phrases (au moins deux).

Par exemple, un verbe au futur dans une première phrase devrait idéalement pouvoir être relié à un verbe au conditionnel présent dans une phrase ultérieure.

Traiter la position des indices au sein des phrases

Le traitement de la position des indices au sein des phrases continue à poser problème et nous n'avons pas trouvé de solution réellement satisfaisante. L'indice positionnel phrastique est traité de la même manière que n'importe quel autre indice englobé ce qui entraîne l'existence potentielle de deux indices différents qui finalement réfèrent à la même entité.

Ainsi, si une phrase contient un adverbial temporel en position initiale, alors elle sera caractérisée par deux attributs distincts : un qui indique la présence d'un adverbial temporel de type x, et un second la présence d'un adverbial temporel en position initiale (dans nos traitements, le type x est volontairement omis).

Nous sommes consciente du biais interprétatif que cette solution entraîne : il est très vraisemblable que les analyses statistiques nous « apprennent » qu'un adverbial temporel de type *ponctuel* est corrélé à la présence d'un introducteur de cadre temporel.

Proposition d'évolution du modèle

Nous envisageons une évolution du modèle permettant de considérer des unités discursives comme les introducteurs de cadre (IC) comme des indices englobants. Il s'agit ici de répondre à la limite présentée dans la section précédente (sur le traitement de la position des indices au sein des phrases).

Une première modification devrait être apportée à l'outil ALIDIS : il faudrait un analyseur de cadres capable de repérer la portée des IC. Ceci permettrait de créer un autre type de segment textuel qui se superposerait aux unités englobantes

un peu à l'image des paragraphes. Cette solution n'a pas pu être développée principalement par manque de temps et parce que nous ne disposons pas d'analyseur de portée pour les introducteurs de cadres[4].

6.3 Format des données en sortie de l'outil OCAS

L'outil OCAS produit en sortie une base de données relationnelle qui sert de référentiel. Nous avons déjà insisté sur le fait que, selon les besoins, les données de sortie peuvent prendre, par exemple, la forme d'un fichier CSV, RTF, XML, EXCEL ou accéder directement à la base de données.

Dans tous les cas, c'est à partir de cette base de données que sont générés des fichiers plats[5] spécifiques à chaque traitement statistique.

D'une manière générale, les fichiers pour les statistiques sont construits de la manière suivante : un enregistrement par individu, soit une ligne pour chacune des 9916 phrases du corpus. Les variables sont créées à partir des indices linguistiques présents dans le corpus.

Sur la base du modèle conceptuel de données, toute unité d'analyse non rectrice d'une autre unité d'analyse devient un individu. Les unités d'analyse rectrices ainsi que les indices discursifs (englobants et englobés) deviennent les variables qui caractérisent les individus.

6.3.1 Choix des individus

Dans un premier temps et de manière relativement intuitive, nous avons opté pour la solution consistant à traiter les indices intra-phrastiques repérés par ALIDIS comme des individus. L'idée consistait alors à spécifier pour chacune d'elles leur appartenance ou non à un segment d'obsolescence. Plusieurs raisons nous ont amenés à rejeter ce choix.

Cette solution rend difficile la gestion d'indices aussi divers que ceux que nous traitons et notamment leur caractère imbriqué et le fait qu'ils peuvent se chevaucher. Il est ainsi difficile de considérer comme des individus une expression temporelle locale au même niveau qu'un paragraphe ou qu'une section ou encore la rubrique des textes.

De plus, notre objectif visant la recherche de combinaisons d'indices, il faut que l'unité d'analyse contienne suffisamment d'indices pour que l'on soit en mesure d'établir des associations entre eux.

Dans un second temps, nous avons pensé à considérer l'unité paragraphe ou, à l'image du TextTiling (Hearst, 1994), une fenêtre glissante dont la taille serait fixée arbitrairement.

[4] Bilhaut (2006) a développé un analyseur de cadre temporel qui pourrait être utilisé dans cette optique.

[5] Dans notre cas, un fichier plat est un fichier dans lequel chaque ligne contient un individu avec toutes les informations que l'on souhaite à son sujet ; tous les individus ont exactement les mêmes informations dans le même ordre.

Le problème principal de cette solution est que le paragraphe n'est pas forcément à mettre à jour entièrement. Très souvent les mises à jour sont plus locales, du niveau de la phrase ou inférieures. De plus, au contact des entreprises d'édition, nous avons remarqué que la pertinence fonctionnelle du paragraphe pouvait parfois être mise en doute : il est fréquent que pour des raisons de mise en page uniquement, le packager[6] impose un saut de ligne que le rédacteur n'avait pas forcément indiqué.

Finalement, la solution choisie considère la phrase comme unité d'analyse régie (*i.e.* comme individu). Tout d'abord, elle présente l'avantage d'être en adéquation avec les contraintes du protocole d'annotation manuelle des corpus (*cf.* p. 46). De plus, la phrase est une unité d'analyse relativement bien définie et plutôt consensuelle en linguistique.

Une phrase peut être incluse dans un indice à grande granularité : par exemple, se trouver dans un paragraphe à l'initiale d'une section, être une phrase introductrice d'un paragraphe. Elle hérite des valeurs indiciaires[7] des segments qui la contiennent.

Parallèlement, une phrase peut inclure des indices plus petits, les indices intra-positionnels et positionnels phrastiques (début, fin de phrase).

Le titre est également considéré comme une unité d'analyse (rectrice) parce qu'il est susceptible, lui aussi, de contenir des indices intra-phrastiques. C'est malgré tout une unité particulière qui présente la particularité de régir une ou plusieurs autres unités d'analyse (cf. schéma 6.2, p. 163). Ces informations sont transformées en variables.

6.3.2 Choix des variables

Les variables sont créées à partir des indices discursifs présents dans le corpus : elles sont constituées du nom générique de l'indice auquel est ajouté sa structure de trait. Ainsi un indice *x* avec la structure de traits $x \begin{bmatrix} A : B \\ C : D \end{bmatrix}$ est transformée en variable $x.A : B; C : D$.

Un indice *y* avec $y \begin{bmatrix} A : B \\ C : E \end{bmatrix}$ devient la variable $y.A : B; C : E$.

Le point sert à délimiter le nom générique de l'indice des élements de la structure de traits ; le point virgule sépare les différents traits de la structure de traits et les deux-points séparent le type du trait (A et C) et sa valeur (B, D ou E).

Par exemple, pour les indices de temps, le nom de la classe générique de l'indice est $exprTemp$ et la structure de traits peut être :

exprTemp $\begin{bmatrix} nature : ponctuel \\ sitTps : coincidence \end{bmatrix}$

Ce qui générera le nom de variable suivant :
$exprTemp.nature : ponctuel; sitTps : coincidence.$

[6] C'est la personne qui est chargée de mettre en forme les livres, les encyclopédies, etc.

[7] Cette expression est reprise des travaux de Widlöcher (2008)

Concernant les indices présents dans les titres ; le fonctionnement est sensiblement le même. On ajoute simplement $title->$ devant le nom de la variable. Ainsi un indice temporel de type ponctuel postériorité présent dans un titre est associé au nom de variable suivant :
$title-> exprTemp.nature : ponctuel; sitTps : posteriorite$

Nous traitons ainsi 146 variables[8] (*i.e.* indices) différentes.

Cette méthode nous permet notamment d'indiquer pour chaque individu le nombre réel d'occurrences de la variable (indice) qui lui est liée, et donc de rester au plus près de la réalité des données du corpus.

L'information d'obsolescence est considérée comme n'importe quel indice discursif. En l'occurrence et parce que l'unité d'analyse centrale est la phrase, il s'agit d'un indice de type englobant.

À partir de la construction des variables sur la base des indices discursifs présents dans le corpus, nous produisons deux vues différentes sur l'*abstraction sémantique* des données. La vue *sorting* est créée pour les statistiques de base et l'ACP et la vue *dataAnalysis* pour l'apprentissage automatique (AA).

Vue *dataAnalysis*

Pour les statistiques de base et l'ACP, nous travaillons sur un fichier constitué de données quantitatives brutes : il s'agit du nombre d'occurrences de l'indice effectivement trouvées dans une même unité d'analyse.

Concernant l'information d'obsolescence, c'est le nombre d'annotateurs qui ont jugé la phrase obsolescente qui est indiqué.

Les valeurs possibles pour une variable vont de 0 à 75.

Le tableau 6.1 (p. 171) illustre le format des données mis en œuvre dans le fichier *dataAnalysis*[9].

Vue *sorting*

Pour l'apprentissage automatique, les données sont discrétisées et typées booléennes. On envisage l'indice en termes de présence *vs* absence.

Le tableau 6.2 (p. 172) illustre le format des données mis en œuvre dans le fichier *sorting*[10].

Exemple

Un exemple concret de transformation des données est présenté dans l'annexe B.2 (p. 257). Sur la base d'un extrait court du corpus Larousse, nous développons les différentes étapes : annotation des indices linguistiques et discursifs avec

[8] Ce nombre correspond au nombre total d'indices effectivement repérés et annotés dans le corpus [ENCYCLO].

[9] Le tableau réel fait 9916 lignes (nombre d'individus) et 146 colonnes (nombre de variables).

[10] Le tableau réel fait 9916 lignes (nombre d'individus) et 146 colonnes (nombre de variables).

Nom des variables	**Type de données/ Nombre d'occurrences**
id	{numérique}
text	*"le contenu de la phrase-individu"*
...	
obsol	{0,4}
...	
V003-encyclopedie.type :GUL	{0,1}
V108-encyclopedie.type :atlas	{0,1}
...	
V116-zone.rubriqueName :ArtLitt	{0,1}
...	
V002-descPhraseType.type :assertion	{0,1}
...	
V034-exprTemps.nature :deictique ;sitTps :coincidence	{0,3}
V063-exprTemps.nature :anaphorique ;sitTps :indetermine	{0,4}
...	
V010-entiteNom.classe :geopolitique	{0,6}
V128-entiteNom.classe :personne	{0,15}
...	
V008-argum.relation :precision	{0,2}
...	
V013-tpsVbx.temps :futur	{0,4}
V006-tpsVbx.temps :présent	{0,11}
...	
V029-ptVue.type :prevision	{0,2}
V031-ptVue.type :distance	{0,2}
...	
V090-periVbs.accomplissement :deroulement	{0,1}
...	
V052-position.typeIndice :exprTemp ;typePos :IC	{0,1}
...	
V123-title–>exprTemps.nature :deictique ;sitTps :coincidence	{0,1}
V128-title–>entiteNom.classe :personne	{0,3}
...	
V129-title–>niveau :3	{0,1}
V066-premierParag.position :debutDivisionSeul	{0,1}
V055-dernierParag.position :finZone	{0,1}
V001-descPhrasePosition.position :debutParagraphe	{0,1}

TAB. 6.1 - *Le format des données dans dataAnalysis : quelques indices/variables (total : 146).*

Nom des variabless	**Type (discrétisé)**
obsol	{0,1}
...	
encyclopedie.type :GUL	{0,1}
encyclopedie.type :atlas	{0,1}
...	
zone.rubriqueName :ArtLitt	{0,1}
...	
descPhraseType.type :assertion	{0,1}
...	
exprTemps.nature :deictique ;sitTps :coincidence	{0,1}
exprTemps.nature :anaphorique ;sitTps :indetermine	{0,1}
...	
entiteNom.classe :geopolitique	{0,1}
entiteNom.classe :personne	{0,1}
...	
argum.relation :precision	{0,1}
...	
tpsVbx.temps :futur	{0,1}
tpsVbx.temps :présent	{0,1}
...	
ptVue.type :prevision	{0,1}
ptVue.type :distance	{0,1}
...	
periVbs.accomplissement :deroulement	{0,1}
...	
position.typeIndice :exprTemp ;typePos :IC	{0,1}
...	
title–>advTemps.nature :deictique ;sitTps :coincidence	{0,1}
title–>entiteNom.classe :personne	{0,1}
...	
title–>niveau :3	{0,1}
premierParag.position :debutDivisionSeul	{0,1}
dernierParag.position :finZone	{0,1}
descPhrasePosition.position :debutParagraphe	{0,1}

TAB. 6.2 - *Le format des données dans sorting : quelques indices (total : 146)*

ALIDIS, création de l'abstraction sémantique sous forme de base de données MYSQL, et enfin, les deux vues possibles, *dataAnalysis* et *sorting*.

6.4 Conclusion

L'outil OCAS est basé sur le modèle d'abstraction sémantique que nous venons de présenter. Concrètement, il permet de transformer nos corpus annoté manuellement et automatiquement en une représentation abstraite des textes et des unités discursives le composant.

Le modèle d'abstraction sémantique nous a permis de concevoir et de créer une base de données relationnelle[11] adaptée à notre objectif et à nos besoins. À partir de la base de données, nous avons créé autant de vues que de traitements différents prévus. Ainsi, deux vues distinctes sur les données ont été créées pour l'étape suivante (la recherche automatique des combinaisons d'indices) : la première est destinée aux traitements statistiques de base et à l'ACP, la seconde est créée pour le système d'apprentissage automatique. Ces deux utilisations ne nécessitent pas les mêmes informations et la même organisation des informations.

Nous mettons également en avant l'aspect autonome[12] de l'outil OCAS notamment quant au format de sortie des données : sur la base d'une même représentaion des données, le format de la vue souhaitée et le type de fichier requis sont entièrement adaptables. Dans notre cas, SPAD[13] accède directement à la base de données alors que pour l'étape d'apprentissage automatique, nous utilisons un fichier csv.

Le caractère reproductible de cet outil OCAS pour d'autres tâches est également au centre de nos motivations : aujourd'hui, nous travaillons sur l'obsolescence ; demain, si nous souhaitons par exemple traiter le cas des passages racistes dans des textes du Web (Valette et Grabar, 2004; Vinot *et al.*, 2003) selon la même méthode, il *suffirait* d'adapter les corpus et les annotations manuelles d'un côté et créer (éventuellement) de nouvelles ressources pour la recherche d'indices discursifs centrées sur le racisme (et non sur l'obsolescence)[14].

L'objectif de l'étape suivante est, à partir des données dont nous disposons, de mettre au jour des configurations, des combinaisons d'indices pour une meilleure description de l'obsolescence et susceptibles de permettre le repérage automatique des segments d'obsolescence.

[11] Décrite dans l'annexe B.1, p. 251

[12] Même s'il est quand même nécessaire de respecter le format de fichier d'entrée, soit du XML sous LINGUASTREAM

[13] Logiciel de traitements statistiques utilisé dans ce travail pour les statistiques de base et l'ACP.

[14] Nous supposons d'ailleurs que certain des indices utilisés pour l'obsolescence pourraient être réutilisés pour d'autres recherches.

Chapitre 7

Étape 3 : Outil STAAT (STatistiques et Apprentissage Automatique sur les Textes)

> *« [L'analyse des données] est un outil pour dégager de la gangue des données le pur diamant de la véridique nature. »* J.-P. Benzécri dans Tufféry (2007, p. IV).

Cette partie s'inscrit dans le cadre général de la fouille de données[1] ou *data mining*. Il s'agit d'une discipline relativement récente qui se situe à l'intersection des domaines de la statistique, de l'informatique, des bases de données, de l'intelligence artificielle, ou encore des interfaces homme-machine. Quel que soit le domaine du *fouilleur*, l'objectif d'une fouille de données est le suivant :

> *« Le data-mining, ou fouille de données, est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de (souvent grandes) bases de données, de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données. »* (Tufféry, 2007, p. 4)

Notre objectif est de mettre en place des techniques statistiques pour :
- *(i)* décrire ce qu'est l'obsolescence en confrontant nos hypothèses linguistiques avec la masse des données annotées ;
- *(ii)* utiliser les techniques d'apprentissage automatique afin de mettre au jour des combinaisons d'indices pertinentes pour le repérage automatique de l'obsolescence.

[1] On rencontre également les termes suivants : *Extraction de Connaissance à partir de Données* (ECD), *Extraction de Connaissance à partir de Bases de Données* (ECBD) ou encore, en anglais, *Knowledge Discovery in Database* (KDD).

Pour répondre à ce double objectif, nous avons mis en place, d'un côté une méthode de type descriptif[2] (ACP), et de l'autre une méthode de type prédictif[3] (apprentissage automatique).

Les statistiques de base (section 7.1) et l'Analyse en Composantes Principales (section 7.2) permettent de décrire les segments obsolescents selon les indices linguistiques considérés indépendamment les uns des autres (statistiques de base), et en termes de combinaisons d'indices (ACP). L'ACP nous renvoie également des informations sur le fonctionnement général des indices dans le corpus. Ces traitements permettent également un tri des variables (*i.e.* des indices) selon leur pertinence et leur significativité au sein des segments d'obsolescence.

L'objectif du processus d'apprentissage automatique (section 7.3) mis en place sur nos données est double : faire émerger de nouvelles connaissances sur l'obsolescence en termes de combinaisons d'indices (orientation descriptive) et *finaliser* notre prototype de repérage automatique de l'obsolescence à partir des règles apprises (orientation prédictive).

Découvrir que deux ou plusieurs variables sont fortement corrélées, comme le font les statistiques de base ou l'ACP, constitue une information de type descriptif même s'il est possible de considérer que la valeur de l'une peut prédire la valeur de l'autre. La frontière entre descriptif et prédictif est souvent floue.

Le schéma 7.1 présente l'organisation de l'outil STAAT mis en œuvre pour répondre à ces objectifs.

7.1 Statistiques de base

Les statistiques de base (ainsi que l'Analyse en Composantes Principales) ont été effectuées à l'aide du logiciel SPAD[4]. Pour cette première étape de traitements statistiques, notre but consiste à valider l'intérêt des indices linguistiques et discursifs pris en compte au regard de l'annotation d'obsolescence (variable *obsol*).

Nous cherchons ainsi à évaluer la relation (*i.e.* la corrélation) entre la variable *obsol* et chacune des autres variables. Les données dont nous disposons pour ces

[2]Les techniques *descriptives* (ou exploratoires) cherchent à extraire de tableaux de données souvent grands, des informations compactes et interprétables. Ces informations sont présentes mais cachées par le volume des données. Dans ce type de techniques, on retrouve des approches très variées : les analyses univariées et bivariées (*cf.* procédure DESCO (p. 176) et avec l'analyse factorielle ((*cf.* ACP, p. 184) ou encore le *clustering* (ou *typologie*). Les techniques descriptives se basent prioritairement sur les relations entre des individus.

[3]Les techniques *prédictives* (ou explicatives) visent à identifier des liens forts entre les variables d'un tableau de données. La découverte de ce lien permet d'identifier des relations entre les variables. Cette équation sert à prédire la valeur de nouveaux individus. Parmi les approches les plus connues, citons par exemple la régression linéaire, l'analyse discriminante, les arbres de décision, la classification automatique hiérarchique, les réseaux de neurones, les séparateurs à vaste marge, etc.

[4]`http://www.spadsoft.com/` : la société Coheris-Spad nous a prêté le logiciel SPAD afin de mener à bien ces recherches. Nous tenons à remercier chaleureusement Mr PLEUVRET, directeur du service de *Recherche et Développement* ainsi que Mme LALEVE pour l'attention qu'ils ont porté à ma demande.

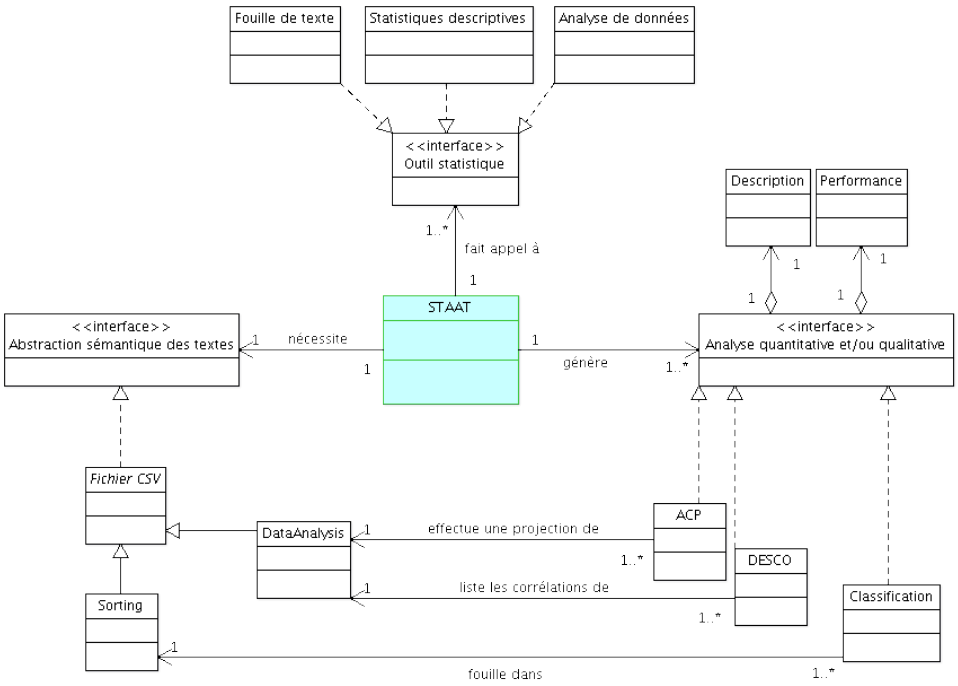


Schéma 7.1 - *L'outil STAAT*

traitements sont des données quantitatives (continues).

SPAD propose des procédures « prêtes à l'emploi » pour effectuer des calculs statistiques sur des données. Nous avons suivi la procédure DESCO.

7.1.1 Méthode : procédure DESCO

Cette procédure permet d'obtenir la caractérisation d'une variable continue en explorant l'ensemble des liaisons qu'elle entretient avec toutes les autres variables continues du fichier. Elle se base sur le cœfficient de corrélation entre deux variables (ou *test de Pearson*, *cf.* documentation du logiciel).

Le cœfficient de corrélation est un indice qui permet de mesurer le degré de dépendance (positive ou négative) entre deux variables : il est représenté par un indice qui varie de 1 (corrélation positive) à -1 (corrélation négative). Plus précisément, un score de :

- 1 signifie que l'indice très fortement associé à l'obsolescence (et jamais à la non-obsolescence) ;
- -1 signifie qu'il est très fortement associé à la non-obsolescence et très rarement

à l'obsolescence[5] ;
- 0 signifie que l'indice est associé indifféremment à des segments obsolescents et non-obsolescents et donc qu'il n'a pas de lien avec la question. Cela indique également que l'indice, du moins utilisé seul, n'est pas utile directement.

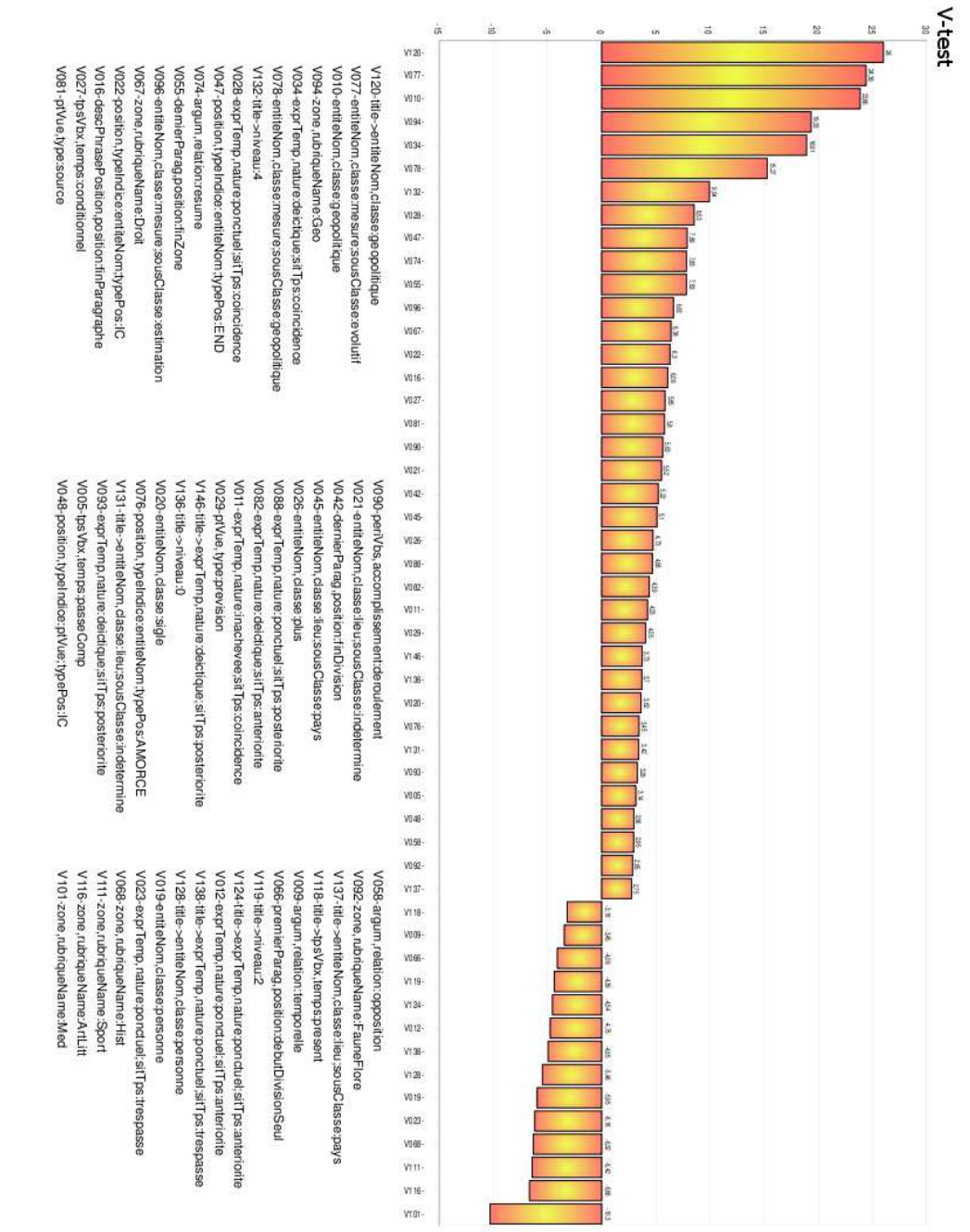
La procédure DESCO propose une estimation des corrélations en termes de *valeurs-tests* (ou *v-test*) : ce test de significativité découle du test de nullité de la corrélation ; plus la valeur-test est élevée, plus l'hypothèse d'une corrélation nulle est facile à rejeter. Les auteurs du logiciel indiquent que si la valeur-test est supérieure à 2, alors le cœfficient est significatif avec un risque d'erreur de 5% . D'une manière générale, plus la valeur-test est grande (en valeur absolue), plus la liaison entre variables est significative et moins le hasard a de chance d'être responsable de celle-ci.

Le graphique 7.2 (p. 179), rend compte des indices qui ont un risque d'erreur de 5 % et une valeur-test supérieure à 2,50 pour les valeurs positives et -2,50 pour les valeurs négatives[6])

Parmi l'ensemble des indices significativement corrélés positivement ou négativement à l'obsolescence (*i.e.* à la variable $obsol$), l'ensemble des types et des niveaux d'indices sont représentés. Il n'y a pas une classe réellement majoritaire ni minoritaire.

[5] C'est également un indice utile et pertinent

[6] Les résultats complets, sous forme de tableau, sont dans l'annexe C.2 (p. 272



Graphique 7.2 - *Les indices linguistiques significativement présents dans les segments d'obsolescence vs. dans les segments non obsolescents*

7.1.2 Résultats et interprétation

La thématique des textes

Tout d'abord, les fiches et entrées encyclopédiques dont la thématique porte sur la Géographie (V094), le Droit (V067) ou la Faune et Flore (V092) semblent plus propices à accueillir des phrases obsolescentes que celles qui parlent de Médecine (V101), d'Art et Littérature (V116), de Sport (V111) ou d'Histoire (V068). Si ces résultats ne sont pas étonnants pour la Géographie, les Arts et Littératures ou encore l'Histoire, ils sont plus surprenant pour la Médecine. Mais dans tous les cas, ce premier constat renforce l'idée qu'il est important de traiter les textes de manière différente selon leur genre.

Le niveau des titres

Concernant le niveau des titres, on observe une opposition entre les individus régis par un titre de niveau 0 (titres d'introduction générale, V136) ou 4 (titres de très bas niveau dans la hiérarchie, plutôt une position conclusive ou illustrative, V132) et ceux régis par un titre de niveau 2 (position intermédiaire dans les textes,V119). Nous avons notamment remarqué que les titres de niveau 4 sont souvent constitués d'un exemple précis, d'une illustration concrète de ce qui a été décrit avant. Les informations sont dans ce cas souvent amenées à évoluer dans le temps.

Les indices de type positionnel

Parmi les indices de type positionnel, s'opposent les derniers paragraphes de section[7] (V042 et V055) aux premiers paragraphes de section (V066). Les positions conclusives semblent donc avoir tendance à accueillir l'obsolescence. La situation est identique pour la position des phrases dans le paragraphe : c'est la dernière phrase des paragraphes (V016) qui est souvent corrélée à l'obsolescence.

Les indices dans les titres

Dans les titres, deux grandes classes d'indices semblent à même de *prédire* la présence de phrases obsolescentes au sein de leur section : les entités nommées et les expressions temporelles. Conformément à nos attentes, les **entités nommées** de type *géopolitique* (« Économie », V120) ou de type *lieu* (sous-types *indéterminé* (« dans certaines régions », V131) et *pays* (« en France », V137)) et les **expressions temporelles** de type *déictique postériorité* (« dans quelques années », V146) apparaissent de manière significative dans des titres qui régissent des segments obsolescents.

À l'opposé, les phrases non obsolescentes sont plutôt régies par des titres contenant des entités nommées de type *personne* (« Omar Khayyam », V128) ou des

[7] Nous ne considérons que les « zones » et les « divisions », soit les niveaux 0 et 1.

expressions temporelles de type *ponctuel antériorité++* (« en 1789 », V138) et *antériorité* (« en mai 68 », V124).

Les indices de type intra-phrastique

Concernant les indices de type intra-phrastique, on observe un certain nombre d'oppositions intéressantes. Tout d'abord, les **entités nommées** semblent jouer un rôle important dans les segments obsolescents. Plus particulièrement, les entités nommées suivantes sont fortement corrélées positivement à l'obsolescence :

- les entités nommées de type *mesure évolutive* (V077) :« 15 % de chômeurs » ;
- de type *géopolitique* (V010) : « population » ;
- de type *mesure géopolitique* (V078) : « 60 millions d'habitants » ;
- de type *mesure estimation* (V096) : « on estime à 3 % » ;
- de type *lieu indéterminé* (V021) : « dans certaines régions » et *lieu pays* : « en France » ;
- de type *superlatif* (V026) : « le pays le plus riche » ;
- et de type *sigle* (V020) : « les OPCVM ».

En revanche, dans les segments non obsolescents (*i.e.* en corrélation négative avec la variable *obsol*), ce sont les entités nommées de type *personne* (« Karl Marx », V019) et de type *lieu pointCardinal* (« dans le nord du pays », V095) qui sont significatives.

De manière assez conforme à nos attentes, les **expressions temporelles** suivantes sont fortement corrélées positivement à l'obsolescence :

- type *déictique* : *coïncidence* (V034) : « aujourd'hui », *antériorité* (V082) : « dans les dernières années », *postériorité* (V093) : « demain » ;
- type *ponctuel coïncidence* (V028) : « en 2007 » ;
- type *ponctuel postériorité* (V088) : « en 2011 » ;
- type *inachevé coïncidence* (V011) : « depuis 2005 ».

À l'opposé, les expressions temporelles référant à une date éloignée du moment de rédaction sont plutôt caractéristiques de la non obsolescence :

- type *ponctuel antériorité* (V023) : « en 1979 » ;
- type *ponctuel antériorité++* (V012) : « en 1936 » ;
- type *déictique anteriorité++* (V030) : « il y a 50 ans » ;
- type *inachevée anteriorité++* (V049) : « depuis les années trente ».

Les **modes verbaux** ont également un impact important. Comme nous le supposions le conditionnel (V027, *cf.* section 3.2.1), qui permet de rapporter des faits tout en exprimant un doute à leur sujet ou d'exprimer qu'on tient une information d'une source non certaine, est en relation positive forte avec l'obsolescence.

Concernant le futur il n'est significatif pour aucun des types de segments, ni dans l'obsolescence, ni dans la non obsolescence. Dans la section 3.2.1 (p. 69) nous avions soulevé la question de considérer le futur comme un temps ou comme un mode. Or il semble important de ne pas le considérer isolément mais au contraire d'observer avec quel(s) autre(s) indice(s) il est corrélé pour être capable de

le classer comme temps ou comme mode : associé au conditionnel, prend-il une valeur modale ? Associé à un temps du passé, est-il de valeur temporelle ?

Concernant les **temps verbaux**, le passé composé (V005) est fortement corrélé positivement à l'obsolescence : au chapitre 3.2 (p. 68) nous avions souligné son rôle comme marque d'énonciation indiquant l'implication du locuteur. À l'opposé, dans les segments non obsolescents, ce sont les temps présent (V006) et passé simple (V040) qui sont représentés.

Le **type des phrases**, contrairement à nos attentes, semble n'avoir aucun impact sur l'obsolescence, du moins envisagé isolément. Il convient cependant de rester prudent sur ce résultat et cette interprétation car il y a en fait très peu de phrases qui sont interrogatives ou exclamatives (99,7 % des phrases du corpus sont assertives).

Les **valeurs aspectuelles** exprimant un procès dans son déroulement (« en cours de », « en train de », V090) sont significatives de l'obsolescence. En revanche, un procès dont l'accomplissement est au début (dans nos traitements, il s'agit exclusivement de la structure « aller + Vinf », V037) est souvent employé dans les situations de futur dans le passé, dans le cadre d'une explication, d'une description historique et donc dans les segments de non obsolescence.

Les **argumentatifs** de type *résumé* (« au total », V074) et *opposition* (« au contraire », « cependant », « en revanche », V058) semblent pertinents dans les segments obsolescents. Ce type d'expression semble particulièrement propice à l'introduction d'exemples, de chiffres de cas illustratifs.

À l'inverse, les argumentatifs de type temporel (« d'abord », « dans un premier temps », « puis », V009) sont significatifs de la non obsolescence. Ce sont des éléments qui, à l'image des MIL (*cf.* section 4.2, p. 99) ont une forte capacité à structurer les textes, à les organiser. Nous avions d'ailleurs souligné leur faible rôle sémantique et leur fort impact structurant.

Enfin, les **expressions de point de vue** de type *source* (« selon l'INSEE », V081) et *prévision* (« un avenir incertain », V029) initient souvent des segments obsolescents. Comme les argumentatifs, ils introduisent souvent des exemples chiffrés spécifiques et illustratifs.

La position des indices au sein de la phrase

Sur la position des indices au sein de la phrase, nous ne pouvons faire aucun constat pertinent : les entités nommées qu'elles soient en début de phrase, en fin de phrase ou en position d'amorce sont toujours significativement liées à l'obsolescence. Les expressions de point de vue le sont également en début de phrase et en position d'amorce. Globalement la position initiale semble jouer un rôle plus marqué que les expressions situées en fin de phrase mais les statistiques ne montrent pas des résultats vraiment marquant allant dans ce sens.

7.1.3 Apports pour le repérage des segments d'obsolescence

Ces statistiques nous permettent de mieux comprendre l'obsolescence et la relation qu'elle entretient avec les autres indices linguistiques et discursifs que nous repérons. Elle montrent que les indices que nous exploitons sont pertinents pour la description et le repérage de l'obsolescence (qu'ils soient corrélés positivement ou négativement à la variable $obsol$).

Ces mesures ont également permis une première description des segments textuels dans lesquels l'obsolescence est susceptible d'apparaître.

Nous avons mené l'expérience suivante : les variables corrélées positivement à l'obsolescence et dont la valeur-test est supérieure à 2^8 sont projetées sur les phrases du corpus ; on mesure ensuite, parmi les phrases qui contiennent au moins l'une de ces variables, combien sont obsolescentes[9]. Les résultats apparaissent dans le tableau 7.1 (p. 183).

La première remarque intéressante est que les indices fortement corrélés à l'obsolescence ne sont pas suffisamment précis : si tous les indices corrélés positivement sont utilisés, soit 51 variables, alors ils repèrent quasiment l'ensemble du texte (8059 sur 9916, soit une précision de 18 %) ; si les indices positionnels (phrastiques, de paragraphe et de document, soit 15 variables) ou les titres (6 variables) sont supprimés, le taux de précision est sensiblement meilleur (23 % dans les deux cas), mais toujours inutilisable concrètement (*i.e.* pour espérer repérer les phrases obsolescentes de manière automatique avec ces indices-là).

Le taux de rappel est naturellement élevé puisque, dans le cas où tous les indices significatifs sont pris en compte, ce sont $8/10^e$ des phrases du corpus qui sont repérées...

	Tous les indices significativement corrélés positivement à *obsol*	Indices de types intra-phrastique + titres	Indices de type intra-phrastique uniquement
Nombre de phrases repérées (Total du corpus : 9916)	8059	5585	5310
Nombre de segments obsolescents repérés (Total dans le corpus : 1508)	1444	1280	1206
Précision	18 %	23 %	23 %
Rappel	95 %	85 %	80 %
F-score	30.3 %	36.2 %	35.7 %

TAB. 7.1 - *Évaluation d'une prise en compte des indices isolés pour le repérage de l'obsolescence*

Exploités de façon indépendante les uns des autres, les 51 indices les plus cor-

[8] Valeur recommandée par le logiciel SPAD.

[9] Les données utilisées sont dans l'annexe C.2, p. 272

rélés positivement à la variable *obsol* ne permettent pas de repérer efficacement les segments obsolescents. Comme nous l'avons déjà souligné, nous faisons l'hypothèse que c'est en termes de configurations, de combinaisons d'indices qu'ils deviendront de bons marqueurs de l'obsolescence. La nature même de certains, notamment les indices positionnels (phrastiques, de paragraphe ou de document), sont d'ailleurs typiquement des éléments qui ne peuvent pas fonctionner seuls dans le cadre d'un repérage automatique de l'obsolescence.

L'étape suivante consiste en la mise en place d'une analyse de données : l'objectif principal est de faire émerger des corrélations complexes entre les indices.

7.2 Analyse de données

Une Analyse en Composantes Principales (ACP) est généralement utilisée pour répondre à l'un des trois objectifs suivants :

1 - un problème met en jeu diverses variables quantitatives (continues) mesurées sur un grand nombre d'individus dont on souhaite tirer des enseignements ;
2 - une ACP peut servir comme étape intermédiaire de calcul avant une analyse ultérieure (régression, discrimination, classification) ;
3 - enfin, elle peut également servir comme technique de compression des données.

Nous visons le premier de ces objectifs : nous cherchons à rendre compte des relations entre les variables également associées à la variable *obsol*. En d'autres termes, l'ACP produit sur des axes les groupes de variables corrélées qui expliquent (le mieux) un phénomène particulier, pour nous, l'obsolescence.

7.2.1 Méthode : Analyse en Composantes Principales (ACP)

Procédure ACP

Une méthode factorielle établit des représentations synthétiques de vastes tableaux de données. L'ACP cherche l'ensemble des corrélations existant entre toutes les variables. L'information composée des corrélations négatives et positives entre les variables est représentée sur des axes (ou facteurs).

Nous appliquons la procédure ACP du logiciel SPAD. L'analyse menée est normée, c'est-à-dire qu'on ne cherche pas à donner plus d'importance aux phrases qui contiennent beaucoup d'indices. Toutes les variables quantitatives (continues) sont actives, soit 146 variables. La variable *obsol* est traitée comme une variable *active* au même titre que toutes les autres variables car ce sont l'ensemble des relations (corrélations) entre *obsol et* les autres variables qui nous préoccupent.

Sur les individus, nous avons appliqué un filtre sur six d'entre eux car ils présentent des caractéristiques extrêmes (*i.e.* la valeur d'une des variables est supérieure à 20). Le nombre d'individus traité est donc ramené à 9910.

Premiers résultats

Les différents résultats sont reproduits dans l'annexe C.3 (p. 277)[10] : l'histogramme des valeurs-propres[11], les différences troisièmes et différences secondes et les intervalles laplaciens d'Anderson[12].

Si l'on observe l'histogramme des valeurs-propres (*cf.* annexe C.3.1, p. 277), on constate qu'il n'y a pas un nombre réduit d'axes représentant significativement toutes les données. Les valeurs propres sont au contraire relativement homogènes. Le pourcentage d'information pour le premier axe est seulement de 2,29 % et de 1,38 % pour l'axe 8. Ces premiers résultats montrent que l'obsolescence est un phénomène qui ne se résume pas par quelques axes seulement et que sa complexité est telle qu'on ne peut pas l'appréhender simplement, *i.e.* avec un nombre réduit d'indices (de variables, d'axes).

Mais notre objectif n'est pas non plus de réduire le nombre des variables à partir des premiers axes pour caractériser l'ensemble des données. Nous visons un but descriptif : rendre compte des relations entre les variables, et plus précisément, là où la variable *obsol* est également présente.

Sélection des axes

Le premier critère de tri des axes consiste à sélectionner les axes au sein desquels la variable *obsol* apparaît : c'est le cas dans 17 axes. Il est intéressant de remarquer que dans les 8 premiers, 6 axes contiennent déjà cette variable[13].

Puis, plus traditionnellement, nous nous basons sur l'histogramme des valeurs propres (*cf.* annexe C.3.1, p. 277) et les tableaux de recherches de paliers (*cf.* annexe C.3.2 et annexe C.3.3 p. 280) pour retenir le nombre optimal de composantes principales. Le principe consiste en la recherche d'un *coude* (ou critère de *Cattell*). Nous avons choisi une coupure entre les axes 8 et 9[14].

Sur la base de cette découpe (les huit premiers axes ou facteurs), la procédure DEFAC de SPAD propose une aide à l'interprétation des facteurs issus de l'ACP. Les items (variables et/ou individus) statistiquement caractéristiques sont sélectionnés et rangés en fonction du critère de la valeur-test que nous avons définie comme devant être supérieure à 2[15].

---

[10] Pour des raisons de place, la matrice des corrélations et la matrice des coordonnées sont disponibles à l'adresse suivante : `http://marion.laignelet.free.fr`.

[11] La valeur-propre représente la qualité d'une information représentée par un axe. C'est une conjonction de variables réunies sur un même axe (*i.e.* une composante principale).

[12] Les intervalles laplaciens permettent de comparer les recouvrements éventuels des axes les uns avec les autres.

[13] Parmi les 20 variables les plus fortes négativement et positivement.

[14] Pour les raisons suivantes : dans l'histogramme des valeurs propres, le pourcentage cumulé est de 12,76 % ; dans le tableau des différences troisièmes, la valeur du palier est de 41,81 et correspond au 4ème palier ; dans le tableau des différences secondes, la valeur du palier est de 58,07 et correspond au 4ème palier également.

[15] La valeur 2 est préconisée par le logiciel : *« Plus la valeur-test est élevée plus la relation linéaire est forte. Une valeur-test inférieure à 2 (en valeur absolue) indique qu'il n'y a pas de liaison linéaire*

Dans l'annexe C.4 (p. 283), nous reproduisons les résultats sous forme de tableaux pour les huit premiers axes. La variable *obsol* est présente dans les axes 2 à 7 si l'on considère les 20 variables dont les coordonnées sont les plus élevées positivement et les 20 variables dont les coordonnées sont les plus élevées négativement.

Interprétation

Cette section se découpe en deux parties : tout d'abord nous présentons des observations d'ordre général sur notre corpus (axes 1 et 8 où la variable *obsol* n'apparaît pas de manière significative) ; ensuite, nous proposons une description de l'obsolescence à travers des combinaisons d'indices émergeant des axes 2 à 7 de l'ACP.

Chaque axe est illustré par des graphiques dans lesquels nous faisons le choix, pour des raisons de visibilité de ne montrer que les codes des variables[16]. Chaque variable et son code associé sont repris dans les explications, ce qui permet une meilleure compréhension[17]. Enfin, pour les axes 2 à 7, les nuages de points (représentant les individus caractéristiques de l'axe) sont associés à la description car ils donnent une idée de la corrélation de la variable obsolescence avec les autres variables[18].

Les résultats détaillés dans la section suivante doivent être interprétés avec précaution. En effet, les résultats ne sont affirmés que pour les valeurs les plus fortes (et les plus faibles) des coordonnées des variables. Plus une coordonnée est proche de 0, moins l'axe correspondant est significatif (la variable ou l'individu participe de moins en moins à la structure mise en évidence par l'axe). Les variables que nous exploitons ont, dans la mesure du possible, des coordonnées sur l'axe au moins supérieures à O,22 (et -O,22), ce qui peut paraître faible.

De plus, les pourcentages cumulés des premiers axes de cette ACP sont plutôt faibles, ce qui nous encourage à rester prudente quant à la fiabilité des résultats.

7.2.2 Résultats et interprétation

Si les résultats de l'ACP nous permettent de mieux comprendre les segments d'obsolescence, ils fournissent également quelques éléments de description du corpus [ENCYCLO] dans sa globalité. Nous ne résistons pas à rendre compte de phénomènes linguistiques (discursifs et structurels) intéressants qui émergent de cette analyse même s'ils ne sont pas directement liés à notre préoccupation principale.

---

entre les variables. » (*in* documentation du logiciel SPAD).

[16] Les correspondances noms de variables/codes se trouvent dans l'annexe C.1 (p. 269).

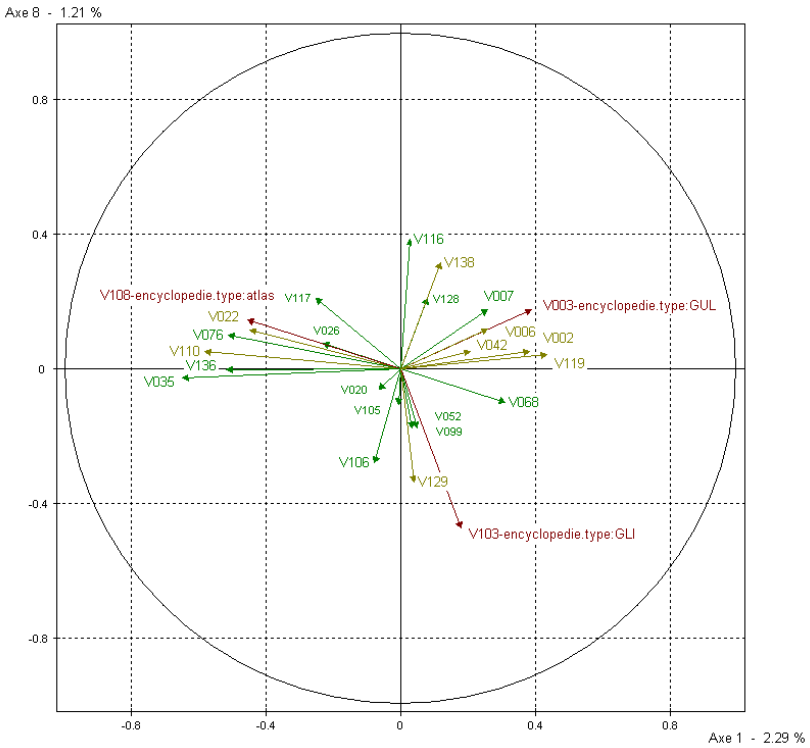
[17] Les variables colorées en vert sont reprises dans les explications ; les variables en jaune sont significatives mais nous n'en faisons pas mention ; les variables en bleu (*obsol*) ou rouge (les sous-corpus) servent à identifier des tendances fortes.

[18] Les points en bleu représentent les individus obsolescents, les points en vert sont les individus non obsolescents.

Remarques générales sur le corpus [ENCYCLO] (facteurs 1 et 8 de l'ACP)

Lors de la phase de constitution du corpus [ENCYCLO] (*cf.* chapitre 2, p. 35), nous avons supposé l'existence d'un *type* encyclopédique défini selon un usage social : en effet, l'objectif d'une encyclopédie est d'informer un lecteur, d'expliquer des faits scientifiques, de décrire des faits historiques, politiques ou sociaux, des états géographiques et économiques, etc. Le type encyclopédique ne se définit donc pas selon une caractérisation linguistique précise mais découle d'un besoin extérieur : toute encyclopédie traitera de géographie, d'histoire, de sciences, d'art, de littérature, de faits de société ou encore de découvertes en médecine. C'est en cela que nous avons considéré le corpus [ENCYCLO] comme homogène et représentatif : les sous-corpus [ATLAS] et [LAROUSSE] sont composés de textes variés.

Caractérisation des sous-corpus par les indices typo-dispositionnels. Nous constatons dès le premier axe factoriel calculé par l'ACP, une opposition marquée entre le sous-corpus [ATLAS] et les sous-corpus [GUL] et [GLI] (dans une moindre mesure pour [GLI]).



Graphique 7.3 - *Vecteurs des variables sur les axes 1 et 8*

Comme nous pouvons le voir sur le graphique 7.3 (p. 187), le sous-corpus [ATLAS] (V108) est associé aux configurations structurelles de type *amorce* (V076)

et aux indices de type *superlatif* (V026) (*cf.* exemple 7.4).

Cette typo-disposition particulière est typique des fiches éditées par les Éditions Atlas : elles sont constituées, à droite, d'un encadré récapitulatif des informations importantes à retenir (*cf.* la figure 2.1, p. 38).

On remarque également que, dans le sous-corpus [ATLAS], les titres de niveau 0 (V117) et 1 (V136) et les paragraphes ne contenant qu'une seule phrase (V035) lui sont également fortement corrélées.

Les exemples 7.4 et 7.5 représentent les individus-types calculés par l'ACP[19] qui ont la plus forte coordonnée sur l'axe 1.

La ville la plus grande : Istanbul, 6,7 millions d'habitants [...]

n° d'individu : 1238262902106

Exemple 7.4 - *Individu caractéristique du sous-corpus* [ATLAS]

Aussi distingue-t-on un capitalisme commercial qui se met en place vers la fin du Moyen Âge et qui perdurera jusqu'au milieu du XVIIIe s. ; un capitalisme industriel et bancaire qui couvre le XIXe s. ; enfin le capitalisme contemporain, ou néocapitalisme, qui apparaît vers la fin du XIXe s., mais qui se développera surtout après la Seconde Guerre mondiale. [...]

n° d'individu : 1237456509848

Exemple 7.5 - *Individu caractéristique du sous-corpus* [LAROUSSE]

Ces observations mettent en évidence des tendances éditoriales à même d'influer sur la forme linguistique et surtout structurelle des textes[20].

Enfin, la sur-représentation des textes traitant d'Économie (V007) et d'Histoire (V068) dans le corpus [LAROUSSE] est également mise en évidence dans cet axe[21].

Caractérisation du sous-corpus [LAROUSSE] par les types d'indices. Le facteur 8 met en opposition les sous-corpus [GLI] et [GUL] (textes issus des Éditions Larousse) sur la base de corrélations d'indices intra-phrastiques et positionnels (*cf.* graphique 7.3, p. 187).

Associé au sous-corpus [GLI], on observe une relation entre les positions d'amorce ou de début de phrase (V106, V052) et les expressions temporelles de type *durée antériorité++* (V105) ou, dans une moindre mesure, les entités nommées de type *sigle* (V020). L'exemple 7.6 rend compte d'un individus caractéristique du sous-corpus [GLI].

[19] Chaque individu est numéroté dans l'ACP ce qui permet de le localiser facilement dans la base de données. Nous indiquons pour chaque exemple cet identifiant unique.

[20] À l'extrême, nous nous demandons s'il serait possible de retrouver l'éditeur d'un texte selon ses formes structurelles et linguistiques.

[21] L'axe 2 de l'ACP met également en avant l'opposition entre Médecine (V101) dans [ATLAS] et Économie (V007) et Histoire (V068) dans [LAROUSSE].

Au XVIe s., l'Espagne devient la puissance prépondérante en Europe. [...]

n° d'individu : 1237458232052

Exemple 7.6 - *Individu caractéristique du sous-corpus* [GLI]

De l'autre côté de l'axe, au sous-corpus [GUL] (V003) sont associées les expressions temporelles de type *antériorité++* en position normale[22] (V138) ou dans un titre (V128) avec les entités nommées de type *personne* (V128) (*cf.* exemple 7.7). De plus, cette combinaison d'indices semble propice dans des textes appartenant à la rubrique Art et Littératures (V116).

Théologien optimiste, convaincu que l'humanité pouvait s'exempter du péché par la connaissance, **Marsile Ficin** affirma que les philosophies païennes recherchaient la Vérité. [...]

n° d'individu : 1238262902712

Exemple 7.7 - *Individu caractéristique du sous-corpus* [GUL]

L'importance des rubriques des textes. Pour poursuivre la description du corpus encyclopédique, nous observons que dans les 15 premiers facteurs de l'ACP, des oppositions systématiques sont établies en fonction des différentes rubriques : Économie/Histoire/Géographie sont opposés à Médecine dans le facteur 1 ; Géographie s'oppose à Histoire dans le facteur 2 ; Économie à Sport dans le facteur 3, Sciences et Techniques à Géographie dans le facteur 7, Sport à Histoire/Art et Littératures dans le facteur 9, etc.

Ces oppositions montrent l'importance à accorder à un traitement qui tienne compte du domaine de chacune des fiches ou entrées de l'encyclopédie. Notre intuition quant au fait de traiter différemment des textes issus de thématiques différentes se confirme (*cf.* chapitre 2.4.1, p. 54).

Remarques particulières sur les segments d'obsolescence (facteurs 2 à 7)

Des combinaisons d'indices temporels et d'entités nommées pour l'obsolescence. Les axes 2 et 3 (graphiques 7.11 et 7.12, p. 192) tendent vers les mêmes conclusions : les entités nommées de type *géopolitique* qu'elles soient en position normale (V010), dans un titre (V120) ou en finale de phrase (V047) sont fortement corrélées à l'obsolescence. L'exemple 7.8 illustre ce cas.

Sur l'axe 2, les entités nommées de type *géopolitique* sont également associée de manière significative aux entités nommées de type *lieu pays* en position normale

[22] Nous parlons de position normale lorsqu'un indice n'est ni en initiale ni en finale. C'est la position par défaut de chaque indice.

Au total, l'**agriculture** n'emploie guère que 1 % des actifs de l'île-de-France, concentrés dans l'**industrie** et de plus en plus dans les **services**, deux secteurs représentés en priorité dans l'**agglomération** de Paris. [...]

n° d'individu : 1237456511712

Exemple 7.8 - *Les entités nommées de type géopolitique dans un segment d'obsolescence (individu représenté sur l'axe 2 - positif)*

(V045) ou dans un titre (V137) et aux entités nommées de type *mesure évolutive* (V077).

Ces observations confirment nos intuitions : il s'avère important de repérer les segments qui contiennent des entités nommées et plus spécifiquement lorsqu'il s'agit d'un nom de pays (peu importe sa position), avec une série de valeurs chiffrées qui se rapportent à un domaine géopolitique prédéfini (taux de population, densité, etc.).

Ces combinaisons tendent à apparaître dans des textes de type Économie (V007), Histoire (V068) ou Géographie (V094).

Sur l'axe 3 et toujours en relation avec l'obsolescence (graphiques 7.11 et 7.12, p. 192), l'association d'indices est sensiblement identique que pour l'axe 2 : les entités nommées de type *géopolitique* sont corrélées aux entités nommées de type *mesure* (*géopolitique* (V078), *évolutif* (V077) et, plus faiblement aux mesures de type *fixe* (V086)). L'exemple 7.9 est représentatif de ce type de combinaisons.

Pays enclavé, dépendant principalement de l'**agriculture** (**85 % de la population active**), des pays limitrophes pour ses débouchés et l'acheminement de ses produits tabac (**63,2 %** des **exportations**), thé (**6,7 %**), canne à sucre (**6,5 %**) et coton (**0,9 %**), le Malawi est classé au **151e rang** en termes de **revenu** national. [...]

n° d'individu : 1238262900553

Exemple 7.9 - *Les entités nommées de type mesure et geopolitique dans un segment d'obsolescence (individu représenté sur l'axe 3 - positif)*

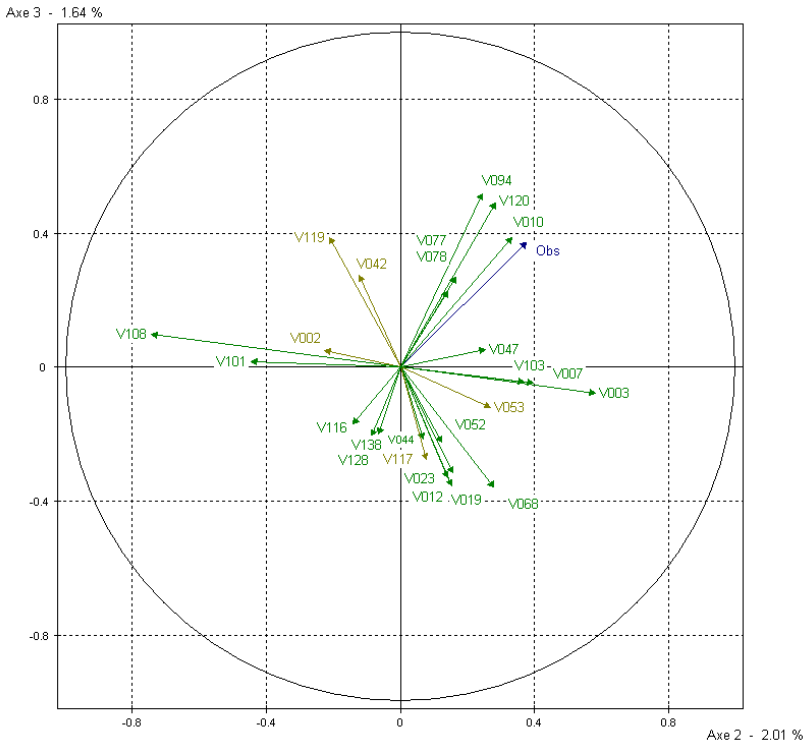
Des combinaisons d'indices temporels et d'entités nommées pour la non-obsolescence. L'axe 3 (graphiques 7.11 et 7.12, p. 192) rend compte également d'une relation intéressante entre des indices temporels et des entités nommées qui sont corrélés négativement à l'obsolescence. Il s'agit des entités nommées de type *personne* (V019) en position normale ou dans un titre (V128) associées à des expressions temporelles de type *ponctuel antériorité++* (V138 et V023) et *antériorité* (V012) dans toutes les positions intra-phrastiques (initiale (V052), finale (V044) et dans un titre (V138)).

Ce type d'association est également corrélé à la rubrique Histoire (V068) ou Arts et Littératures (V116). Ce type de segment est alors peu susceptible de mises à jour comme l'illustre l'exemple 7.10.

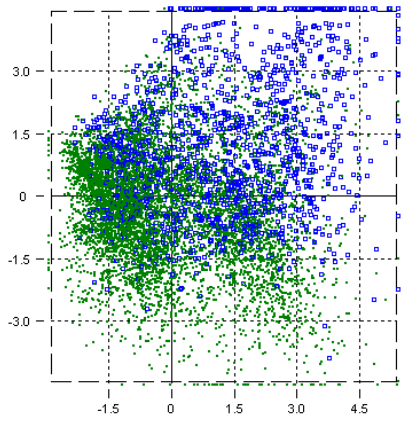
Ses principaux représentants sont le contrebassiste **Charlie Mingus** (1922-1979), les saxophonistes **Julian** « Cannonball » **Adderley** (1928-1975) et **Sonny Rollins** (né **en 1930**), le pianiste **Horace Silver** (né **en 1928**) et le batteur **Art Blakey** (né en 1919), qui fondent les **Jazz Messengers en 1953**, ainsi que toute une pléiade de musiciens exceptionnels : les trompettistes **Clifford Brown** (1930-1956), **Lee Morgan** (1938-1972), le saxophoniste **Johnny Griffin** (né **en 1928**), les pianistes **Wynton Kelly** (1931-1971), **Red Garland** (1923-1984) et **Tommy Flanagan** (1930-2001), l'organiste **Jimmy Smith** (né **en 1925**), le guitariste Wes Montgomery (1928-1968), les batteurs **Roy Haynes** (né **en 1926**) et **Philly Jo Jones** (1923-1985). [. . .]

n° d'individu : 1238262902967

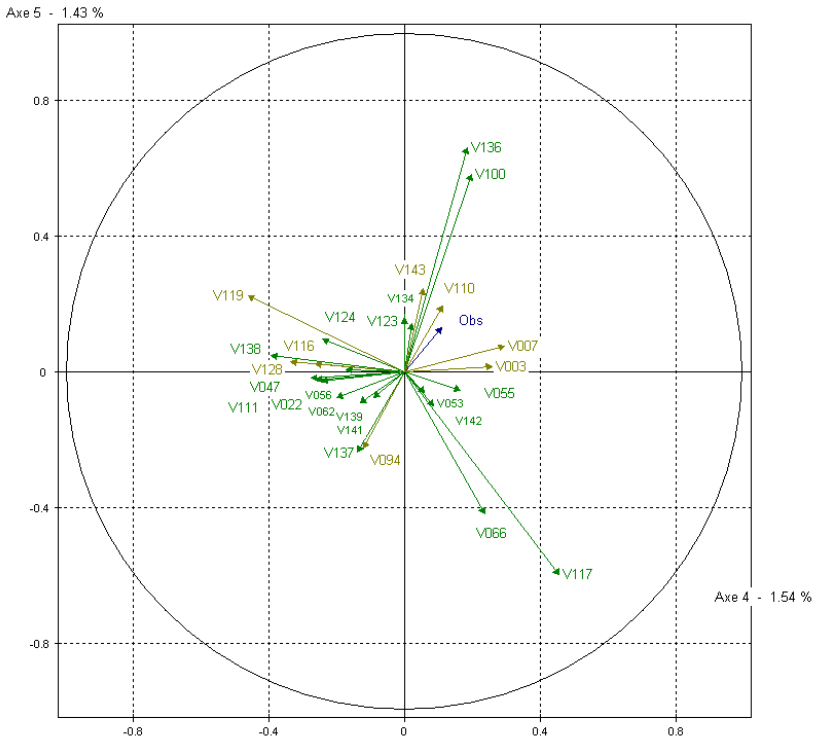
Exemple 7.10 - *Les entités nommées de type personne et le temps de type antériorité dans un segment non obsolescent (individu représenté sur l'axe 3 - négatif)*



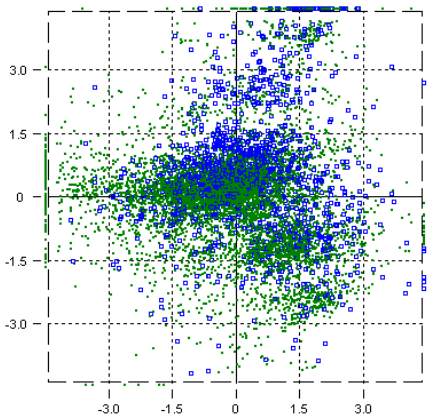
Graphique 7.11 - *Vecteurs des variables sur les axes 2 et 3*



Graphique 7.12 - *Coordonnées des individus sur les axes 2 (abscisse) et 3 (ordonnée)*



Graphique 7.13 - *Vecteurs des variables sur les axes 4 et 5*



Graphique 7.14 - *Coordonnées des individus sur les axes 4 (abscisse) et 5 (ordonnée)*

L'importance de la typo-disposition. L'axe 4 (graphiques 7.13 et 7.14, p. 193) permet de décrire l'obsolescence selon des considérations typo-dispositionnelles : les segments régis par un titre de niveau 1 (V117) et en position d'introduction de section (V066) ou en position de conclusion générale (V055) semblent être des positions privilégiées pour accueillir des segments d'obsolescence.

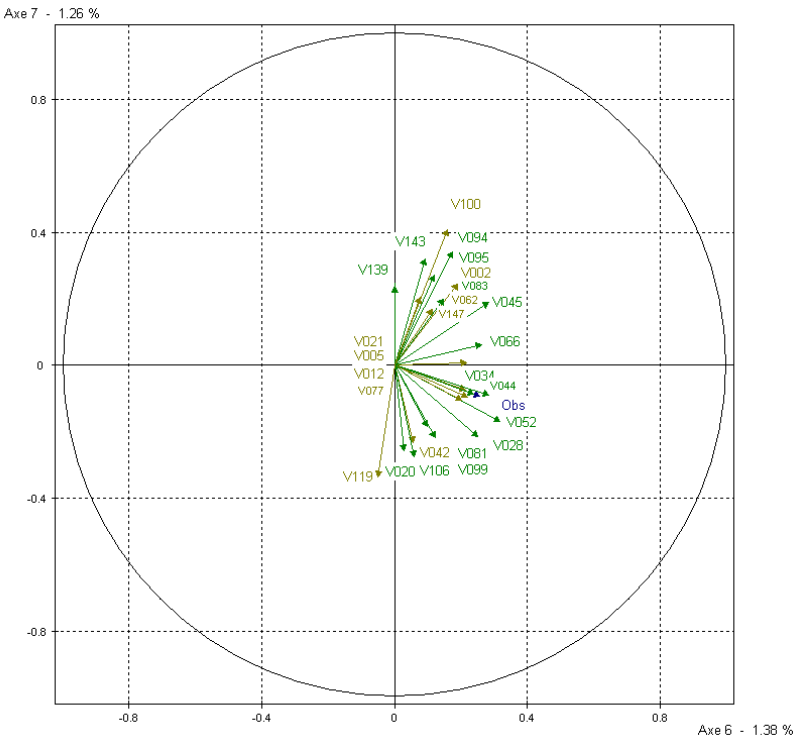
À l'opposé, cet axe nous apprend qu'un segment textuel régi par une rubrique de type Sport (V111), composé d'expressions temporelles de type *ponctuel antériorité++* (V124) et *antériorité* (V138) dans un titre, d'entités nommées de type *personne* (V128), de type *mesure indéterminée* (V056) et de type *lieu ville* (V062) (position phrastique normale, à l'initiale (V022) et en finale (V047)) ne sera très probablement pas obsolescent. Ces observations sont à rapprocher des observations faites sur l'axe 3 (*cf.* p. 190).

L'importance du point de vue temporel du rédacteur. L'axe 5 (graphiques 7.13 et 7.14, p. 193) permet de décrire les individus obsolescents composés d'expressions temporelles de type *déictique coïncidence* (V123) et d'expressions de point de vue de type *prévision* (V134). L'exemple 7.15 illustre ces cas. Ces segments sont également généralement situés en position d'introduction générale (V100, V136).

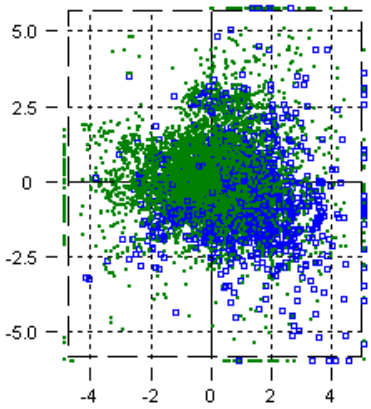
Leur sauvegarde et leur protection **restent** un **défi majeur** posé aux populations mondiales **dans les prochaines années**. [...]

n° d'individu : 1238262902112

Exemple 7.15 - *Les expressions temporelles de type déictique coïncidence et les expression de point de vue de type prévision dans un segment obsolescent (individu représenté sur l'axe 5 - positif)*



Graphique 7.16 - *Vecteurs des variables sur les axes 6 et 7*



Graphique 7.17 - *Coordonnées des individus sur les axes 6 (abscisse) et 7 (ordonnée)*

Des indices intra-phrastiques associées à des indices positionnels textuels. Dans le facteur 6 (graphiques 7.16 et 7.17, p. 195), les segments obsolescents (*cf.* exemple 7.18) sont caractérisés par la présence conjointe d'expressions temporelles de type *coïncidence* (V028, V034) dans les positions normales de phrase, en initiale (V052) et en finale (V044) ainsi que des entités nommées de type *lieu pays* (V045). La position de premier paragraphe de division (V066) semble également déterminante. Ces observations sont à mettre en parallèle avec les observations faites sur l'axe 2. Ici, ce n'est pas le titre qui est corrélé aux indices temporels et aux entités nommées de lieux mais la position textuelle.

Située au sud-est de l'**Europe**, la **Grèce** (en grec, Elláda ou Ellás), officiellement **République hellénique**, (Elliniki Dimokratia) est **aujourd'hui** une démocratie parlementaire, limitée par la **Macédoine** et la **Bulgarie au nord**, l'**Albanie au nord-ouest**, la **Turquie au nord-est**, la **mer Égée à l'est**, la **mer Ionienne à l'ouest** et la **mer Méditerranée au sud**. [...]

n° d'individu : 1238262901032

Exemple 7.18 - *Les expressions temporelles de type coïncidence et les entités nommées de type lieu pays dans un segment obsolescent (individu représenté sur l'axe 6 - négatif)*

Le point de vue du rédacteur : se distancer de ses propos. Le facteur 7 (graphiques 7.16 et 7.17, p. 195) met en évidence les phrases constituées d'expressions temporelles de type *ponctuel coïncidence* (V028) en position normale et d'amorce (V106), d'entités nommées de type *sigle* (V020) et d'expressions du point de vue de type *source* (V081) (*cf.* exemple 7.19). La rubrique Sciences et Techniques (V099) est fortement corrélée à ces variables. Ces phrases ont une probabilité forte d'être obsolescentes.

D'après le BIT, dans les pays de l'OCDE, 37 % des travailleurs étaient affiliés à un syndicat en 1975 ; ils n'étaient plus que 28 % en 1988 et l'érosion des effectifs s'est poursuivie au cours de la décennie suivante (taux aujourd'hui souvent inférieur à 20 %), notamment en lien avec le recul de l'identité ouvrière. [...]

n° d'individu : 1237458232636

Exemple 7.19 - *Les expressions temporelles de type ponctuel coïncidence, les entités nommées de type sigle et l'expression du point de vue de type source dans un segment obsolescent (individu représenté sur l'axe 7 - positif)*

De l'autre côté de l'axe 7, une phrase composée de temps verbaux au passé simple (V143), d'entités nommées de type *lieu pointCardinal* (V095) et de type *lieu ville/rivière/pays* (V139, V083 et V045) ne sera probablement pas à mettre à jour. Ces corrélations semblent fréquemment associée aux phrases issues de textes de Géographie (V094).

7.2.3 Apport de l'ACP pour le repérage des segments d'obsolescence

Les axes émergents de cette ACP sont des objets complexes à décrire du fait du nombre élevé de variables prises en compte (146) et du fait que cette ACP ne compresse rien (seulement 12,79 % cumulé à l'axe 8).

Tout en restant prudent sur l'interprétation des résultats, l'ACP nous a permis d'approfondir la description des variables corrélées à l'obsolescence et, par extension celle des combinaisons d'indices. Les résultats de l'ACP nous encouragent d'ailleurs dans ce sens.

Nous avons expérimenté un petit outil *fait maison* créé sur la base des observations des corrélations entre les variables dans les axes 2 à 7 : nous avons extrait un ensemble de règles de décision sur le caractère obsolescent ou non d'un individu donné. En d'autres termes, nous faisons *à la main* ce qu'un système d'apprentissage automatique est capable de faire.

Huit règles de combinaison d'indices[23] sont produites sur la base des axes 2 à 7. Elles sont projetées sur nos données : le nombre de segments obsolescents qu'elles permettraient de récupérer est alors évalué. Les résultats sont fournis dans le tableau 7.2.

	Règles issues des facteurs 2 à 7
Nombre de phrases repérées (Total du corpus : 9910)	834
Nombre de segments obsolescents repérés (Total dans le corpus : 1508)	377
Précision	45 %
Rappel	25 %
F-score	32,1 %

TAB. 7.2 - *Évaluation des 8 règles créées à partir des axes 2 à 7 de l'ACP*

Les résultats ne sont globalement pas bons. Avec huit règles créées à partir de six axes, un segment obsolescent sur quatre est repéré et sur deux segments obsolescents, un l'est réellement.

Si on compare ces résultats avec ceux de l'expérimentation menée à partir des statistiques de base (*cf.* tableau 7.1 (p. 183), et malgré un F-score relativement aussi mauvais dans tous les cas, on constate que les taux de précision et de rappel sont inversés : dans ce cas, ce n'est pas la *quantité* de segments obsolescent qui est visée mais plutôt la *qualité*, ce qui *a priori* nous intéresse moins.

D'une manière générale, l'ACP montre que le phénomène de l'obsolescence est non trivial et qu'il n'est pas possible de chercher à l'appréhender avec peu d'indices ou des configurations d'indices simples. Aucune tendance forte n'est exprimée. L'expérimentation que nous avons menée (*cf.* tableau 7.2) montre qu'il est difficile de chercher à rendre compte de ce phénomène de manière simple (*i.e.* à la

[23] Elles sont disponibles dans l'annexe C.5 (p. 293).

main). C'est pourquoi nous avons décidé de mener un apprentissage automatique des combinaisons présentes dans les segments d'obsolescence. La section suivante présente le système d'apprentissage automatique basé sur des règles d'association qui a été mis en place[24].

7.3 Apprentissage automatique

Un système d'apprentissage automatique extrait des informations nouvelles à partir de données : il s'agit de découvrir automatiquement des règles cachées au sein de grandes masses d'informations. Ces nouvelles connaissances peuvent alors être transposées sur de nouvelles données afin de permettre la meilleure prise de décision possible.

Dans notre cas, nous cherchons à décrire l'obsolescence et à formuler des règles qui permettront par la suite de repérer automatiquement dans des textes nouveaux (*i.e.* non annotés manuellement) des segments d'obsolescence.

Concernant la technique d'apprentissage automatique mise en oeuvre, nous avons fait appel au savoir et savoir-faire de Francois Rioult[25] (Laignelet et Rioult, 2009), chercheur au GREYC (Caen). Nous avons utilisé le logiciel mvMiner[26] qu'il a développé et qui s'intègre à LINGUASTREAM.

7.3.1 Méthode : les règles d'association

Nous utilisons dans ce travail un système à base de règles d'association : il s'agit d'une procédure prédictive qui vise la description des segments d'obsolescence et la projection des règles apprises sur un corpus de test.

Les règles d'association présentent la particularité d'être bien adaptées à une étude exploratoire car toutes les règles produites sont conservées même si elle sont triviales et/ou redondantes. notre objectif n'est en effet pas de rechercher les règles idéales en supprimant certaines d'entre elles comme dans le cas des arbres de décision par exemple. Toutes les associations sont considérées comme potentiellement intéressantes : dans notre cas, nous filtrons les règles qui concluent sur la classe $obsol$. En d'autres termes, cette méthode permet de ne pas poser d'hypothèses sur le modèle, de ne pas choisir *a priori* les descripteurs. De manière conforme à notre démarche *top-down*, les indices et combinaisons d'indices pertinents émergent de notre corpus.

[24] Le logiciel SPAD propose également une méthode de *classification hiérarchique* qui se base sur les résultats de l'ACP. Nous espérons pouvoir mettre en œuvre cette méthode.

[25] Un grand merci à François Rioult pour toute l'aide qu'il m'a apportée et qui m'a permis d'achever ce travail.

[26] `http://boita.info.unicaen.fr/plone/data-mining/linguastream/fouille-de-donnees-avec-linguastream/`

Les mesures d'association

L'image du *panier de la ménagère* constitue l'exemple le plus connu pour expliquer cette méthode : un système à base de règles d'association recherche les associations entre des objets (dans le cas du panier de la ménagère, on recherche quels produits tendent à être achetés ensemble sur la base des associations entre produits apparaissant sur les tickets de caisse).

En d'autres termes, on recherche les regroupements, les associations potentielles d'objets. Une règle d'association vise donc à prédire une classe moyennant un ensemble de conditions (*i.e.* un *itemset*).

Les systèmes à base de règles d'association suivent généralement les étapes suivantes :

1. génération des itemsets fréquents
2. génération des règles entre ces itemsets
3. évaluation et validation des règles

La prédiction associée à la conclusion de la règle n'est pas limitée à une seule classe d'attribut mais peut être associée à une ou plusieurs combinaisons d'attributs[27]. Un filtre est ensuite appliqué pour la classe recherchée : en ce qui nous concerne, l'obsolescence.

Une règle d'association est de la forme suivante :

$Si\ A, alors\ B$ où

- A et B sont des conjonctions d'attributs
- A est la condition, la prémisse de la règle
- B est la conclusion (sur une valeur de classe).

L'association entre A et B est mesurée par :

- la *fréquence* d'apparition d'une règle (ou *support*) : le nombre de fois où l'association $A \rightarrow B$ est présente, rapportée au nombre de règles contenant A ou B ;
- la *confiance* : le nombre de fois où l'association $A \rightarrow B$ est présente, rapportée au nombre de présences de A.

Pour nos expériences, une adaptation de Li *et al.* (2001) a été implémentée dans MVMINER[28], outil de fouille de données capable de considérer différents types de règles (Rioult *et al.*, 2008) à des fins exploratoires.

Nous avons ainsi testé trois types de règles :

[27] Contrairement aux systèmes de classification comme les règles de décision pour lesquels il faut déterminer *a priori* la classe à traiter.

[28] Développé par François Rioult et disponible sur `http://boita.info.unicaen.fr/plone/author/frioult`. Il est par ailleurs intégrable et intégré à la plateforme LINGUASTREAM

- des **règles d'association classiques** telles que nous venons de les décrire.
- des **règles d'association *généralisées*** (ou disjonctives) : à la différence des règles d'association classiques, elles contiennent des règles positives ou/et négatives tant dans la prémisse que dans la conclusion de la règle.
- des **règles construites sur des *motifs émergents*** : dans ce cas, la valeur de la classe à analyser est précisée. On s'éloigne donc des règles d'association (qui recherchent toutes les règles possibles pour toutes les valeurs de classe possibles) pour entrer dans le domaine des règles de classification à proprement parler (car on indique *a priori* la classe à décrire).

Paramétrage du classifieur

Les classifieurs sont paramétrés de la manière suivante :
- u correspond à la **profondeur** (ou **longueur**) de la règle. Il s'agit d'indiquer le nombre d'attributs pour les règles. Dans notre cas, elle est à 1 (valeur conseillée pour les règles d'association classique).
- $delta$ indique le nombre (ou la fraction) d'**erreurs tolérées**. Nous avons choisi un $delta$ de 5, soit 5 erreurs tolérées par règle.
- $minsup$ évalue le seuil de **support** (entre 0 et 1). Nous avons établi le seuil de support à 0.0001 (soit 1 pour 1000). Cela signifie qu'une règle doit fonctionner au moins sur 9 phrases (puisque le corpus contient 9916 phrases). Ce taux peut paraître faible : il est pourtant nécessaire du fait que certains descripteurs sont peu fréquents ans le corpus d'apprentissage (moins de 20 occurrences). Il est important de les considérer car ils sont susceptibles d'être pertinents pour l'obsolescence. L'obsolescence étant un phénomène rare, il n'est pas étonnant que des descripteurs (ou des combinaisons de descripteurs) rares soient pertinents pour la classe $obsol$.

Une classification supervisée consiste à identifier des classes connues *a priori* et pour lesquelles on dispose d'exemples (des objets que l'on peut décrire selon des traits descriptifs (attributs, caractéristiques, features).

> *« Le but est de trouver une description générale et caractéristique décrivant une classe sans avoir à énumérer tous les exemples de cette classe. Il faut découvrir ce que les exemples ont en commun et qui peut donc être induit comme étant la description de la classe. »* (Loudcher-Rabaseda, 1996, p.11)

Dans ce travail, la classe à définir est la classe *obsolescence* : le corpus [ENCYCLO] a été annoté par des experts qui ont étiqueté chaque individu par une valeur de classe : obsolescent ou non obsolescent. Les individus sont ensuite regroupés selon les classes prédéfinies qui serviront de base d'apprentissage des règles de prédiction. Après génération des règles, nous filtrons celles qui concluent sur la classe $obsol$.

Le format des données diffère sensiblement de celui qui a été mis en place pour les statistiques descriptives (*cf.* section 6.3, p. 168). Pour des raisons d'explosion combinatoire, nous avons été contraint de discrétiser les valeurs quantitatives des variables à une opposition booléenne (0 / 1). Le nombre d'individus à traiter reste le même que précédemment : 9916 phrases à classer ; le nombre d'attributs est de 146.

Nous avons également créé quatre autres vues du corpus d'apprentissage. Ces vues permettent de considérer des variations sur le type des attributs à prendre en compte par le classifieur. Ainsi, les ressources suivantes sont disponibles :

- *corpusComplet* : une vue qui prend en compte tous les indices ;
- *corpusIPseuls* : une vue qui prend en compte uniquement les indices intra-phrastiques ;
- *corpusIPHierar* : une vue qui prend en compte les indices intra-phrastiques et les indices hiérarchiques ;
- *corpusIPPos* : une vue qui prend en compte les indices intra-phrastiques et les indices positionnels ;
- *corpusEpure* : un corpus « épuré » dans lequel sont enlevées les variables non significatives (en fonction des résultats des statistiques de base, DESCO, et de l'ACP).

Nous comparons l'apport des différents indices et niveaux d'indices pour la performance de notre outil de classification des phrases selon leur caractéristique d'obsolescence. Nous présentons ces comparaisons et leurs résultats dans la section 7.3.3 (p. 207).

La section qui suit propose une description qualitative des règles générées par le système d'apprentissage automatique.

7.3.2 Analyse des connaissances obtenues : retour sur la description des segments obsolescents

Les règles d'association générées

Nous représentons dans le schéma 7.20 (p. 202) les 30 premières règles d'association (classiques) (sur environ 1300, redondantes) qui concluent sur la valeur de classe *obsol* en sortie du classifieur. Elle sont classées selon un score de pertinence qui n'est pas reproduit ici. Le taux de couverture des ces règles est de 1 pour 1000 : une règle concerne au moins 9 phrases.

Description des règles d'association concluant sur la classe *obsol*

Nous avons regroupé les règles d'association en sept grandes classes d'indices.

$1-premierParag.position:debutDivision \wedge zone.rubriqueName:NULL \wedge title.entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$2-premierParag.position:debutDivision \wedge encyclopedie.type:GLI \wedge title.entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$3-encyclopedie.type:GLI \wedge title.entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$4-entiteNom.classe:mesure; sousClasse:geopolitique \wedge title.entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$5-entiteNom.classe:mesure; sousClasse:evolutif + title- > entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$6-premierParag.position:debutDivision \wedge title- > entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$7-exprTemp.nature:deictique; sitTps:coincidence \wedge entiteNom.classe:geopolitique; sousClasse:indetermine \wedge zone.rubriqueName:NULL \longrightarrow classe:obsol$
$8-entiteNom.classe:mesure; sousClasse:geopolitique \wedge zone.rubriqueName:Geo \longrightarrow classe:obsol$
$9-exprTemp.nature:deictique; sitTps:coincidence \wedge entiteNom.classe:mesure; sousClasse:evolutif \longrightarrow classe:obsol$
$10-entiteNom.classe:geopolitique; sousClasse:indetermine \wedge encyclopedie.type:GLI \wedge title- > entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$11-exprTemp.nature:deictique; sitTps:coincidence \wedge entiteNom.classe:geopolitique; sousClasse:indetermine \longrightarrow classe:obsol$
$12-exprTemp.nature:deictique; sitTps:coincidence \wedge entiteNom.classe:geopolitique; sousClasse:indetermine \longrightarrow classe:obsol$
$13-entiteNom.classe:lieu; sousClasse:ville \wedge position.typeIndice:entiteNom; typePos:END \wedge title- > entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$14-premierParag.position:debutDivision \wedge entiteNom.classe:geopolitique; sousClasse:indetermine \wedge title- > entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$15-exprTemp.nature:deictique; sitTps:coincidence \wedge encyclopedie.type:GLI \longrightarrow classe:obsol$
$16-descSentencePosition.position:debutParagraph \wedge exprTemp.nature:deictique; sitTps:coincidence \wedge tpsVbx.temps:passeComp \longrightarrow classe:obsol$
$17-entiteNom.classe:mesure; sousClasse:evolutif \wedge entiteNom.classe:geopolitique; sousClasse:indetermine \longrightarrow classe:obsol$
$18-entiteNom.classe:mesure; sousClasse:evolutif \wedge entiteNom.classe:geopolitique; sousClasse:indetermine \longrightarrow classe:obsol$
$19-exprTemp.nature:deictique; sitTps:coincidence \wedge premierParag.position:debutDivisionzone.rubriqueName: \longrightarrow classe:obsol$
$20-entiteNom.classe:mesure; sousClasse:evolutif \wedge tpsVbx.temps:conditionnel \wedge title- > entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$21-descSentencePosition.position:finParagraph \wedge zone.rubriqueName:Geo \wedge title- > entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$22-descSentencePosition.position:debutParagraph \wedge entiteNom.classe:mesure; sousClasse:geopolitique \longrightarrow classe:obsol$
$23-descSentencePosition.position:finParagraph \wedge premierParag.position:debutDivision \wedge title- > entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$24-exprTemp.nature:deictique; sitTps:coincidence \wedge zone.rubriqueName:\wedge title- > entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$25-exprTemp.nature:deictique; sitTps:coincidence \wedge encyclopedie.type:GLI \wedge title- > entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$26-descSentencePosition.position:debutParagraph \wedge exprTemp.nature:deictique; sitTps:coincidence \wedge encyclopedie.type:GLI \longrightarrow classe:obsol$
$27-entiteNom.classe:mesure; sousClasse:evolutif \wedge title- > entiteNom.classe:geopolitique \longrightarrow classe:obsol$
$28-entiteNom.classe:mesure; sousClasse:evolutif \wedge entiteNom.classe:geopolitique; sousClasse:indetermine \wedge zone.rubriqueName:Geo \longrightarrow classe:obsol$
$29-entiteNom.classe:mesure; sousClasse:evolutif \wedge entiteNom.classe:geopolitique; sousClasse:indetermine \wedge zone.rubriqueName:\longrightarrow classe:obsol$
$30-entiteNom.classe:geopolitique; sousClasse:indetermine \wedge entiteNom.classe:mesure; sousClasse:geopolitique \longrightarrow classe:obsol$

Schéma 7.20 - *Exemple de règles qui concluent sur la valeur de classe obsol*

L'association entre des indices hiérarchiques et des indices intra-phrastiques est relativement fréquente (par exemple dans les règles 4, 5, 20, 24, etc.).

Ainsi, les titres comprenant une expression de type géopolitique (« La population ») sont fréquemment associés à un indice de plus bas niveau comme une entité nommée de type géopolitique (« 100 000 hab. »), mesure (« 78 % ») ou lieu (« Madrid », « Barcelone ») ou encore une expression temporelle de type *déictique coïncidence*. La relation est forte également entre des titres contenant un verbe au conditionnel et des phrases dans lesquelles se trouve une entité nommée de type *lieu*.

La population

§ La population s'est urbanisée (près de **78 %** de la population vit en ville). Une quarantaine de villes ont plus de **100 000 hab.**, dominées par les pôles de **Madrid** et **Barcelone**. [...]

Source : Corpus GLI

Exemple 7.21 - *Exemple de combinaisons d'indices hiérarchiques et d'indices intra-phrastiques*

La corrélation entre des indices positionnels textuels et des indices intra-phrastiques est importante (par exemple dans les règles 6, 14, etc.) : le premier paragraphe d'une division associé à une expression temporelle de type *déictique coïncidence* (« aujourd'hui ») ou à une entité nommée de type *géopolitique* entraînent souvent l'obsolescence du segment dans lequel l'indice intra-phrastique apparaît. Il en est de même lorsque qu'une entité nommée de type *mesure évolutive* apparaît dans le dernier paragraphe d'une section (division).

Concernant la position des indices au sein de phrases, la position de fin de phrase est pertinente dans la classe *obsol*, notamment lorsque cette position est occupée par une entité nommée de type *mesure* ou *lieu ville* (règle 13).

Concernant la position des phrases au sein des paragraphes, les premières phrases de paragraphe contiennent souvent des indices temporels de type *déictiques coïncidence* ou des entités nommés de type *mesure évolutive* ou *mesure géopolitique* lorsqu'elles sont obsolescentes (par exemple la règle 22). De plus, lorsqu'un verbe au conditionnel associé à une entité nommée de type *géopolitique* est en fin de paragraphe, alors la mise à jour du segment est fortement prévisible.

§ L'Union européenne à elle seule se **serait** dépossédée d'un patrimoine de **215 milliards de dollars**. [...]

Source : Corpus GLI

Exemple 7.22 - *Exemple de position des phrases au sein des paragraphes*

L'obsolescence est également mise en valeur par l'association marquée entre plusieurs indices intra-phrastiques. Ainsi, une phrase sera obsolescente si une expression temporelle de type *déictique coïncidence* est reliée à une entité nommée de type *géopolitique* ou de type *mesure évolutive* (par exemple dans les règles 7, 9, 11, etc.).

§ Les Noirs, représentent **aujourd'hui 12 %** de la population ; plus de **50 %** d'entre eux sont **encore** concentrés dans le Sud historique. [...]

Source : Corpus GLI

Exemple 7.23 - *Combinaison de plusieurs indices intra-phrastiques - 1*

La présence conjointe d'une entité nommée de type *géopolitique* avec une entité nommée de type *mesure évolutive* ou de type *lieu* est également pertinente pour le repérage des segments obsolescents (dans les règles 9, 11 ou encore 20).

§ Les **industries** de pointe (**11 %** des emplois salariés dans les activités high-tech) ont bien représentées à **Lyon** et à **Grenoble** (électronique, micro- et nanotechnologies). [...]

Source : Corpus GLI

Exemple 7.24 - *Combinaison de plusieurs indices intra-phrastiques - 2*

Des règles mettant en relation trois niveaux différents d'indices sont apprises par le système.

Ainsi, selon la règle 23, une phrase sera considérée comme obsolescente si *(i)* elle est la dernière du paragraphe, si *(ii)* le paragraphe est en première position dans la division ou en fin de division et si *(iii)* le titre de la section contient une entité nommée de type *géopolitique*.

Il en est de même pour une phrase *(i)* contenant une entité nommée de type *lieu ville*, (ii) qui est située en dernière position de paragraphe et *(iii)* le titre chapeautant la section contient une entité nommée de type *géopolitique* (par exemple dans les règles 21 ou 23).

Les résultats montrent enfin qu'un indice externe comme le domaine ou la rubrique peut jouer un rôle important.

Par exemple, la règle (21) suivant laquelle une phrase est obsolescente si elle est la dernière du paragraphe et que le titre qui la régi contient une entité nommée de type *géopolitique* est validé pour la rubrique Géographie.

En revanche la règle (23) stipulant qu'une phrase doit être en première position de paragraphe, que le paragraphe doit être le premier de la section et que le titre qui la régit contient une entité nommée de type *géopolitique* s'applique dans tous les cas, sans distinction du domaine.

Pour conclure, les règles décrites sont massivement orientées vers le repérage des informations géographiques. La classe *géopolitique* que nous avons créée en fonction des textes du corpus est très présente dans les segments d'obsolescence. Les aspects temporels (*coïncidence* et *déictique* sont également très importants dans ces segments. Mais il est important de remarquer que ces règles ne s'appliquent pas uniquement aux textes appartenant au domaine de la Géographie (alors que c'est sur leur base que le lexique *géopolitique* a été créé (*cf.* section 3.5.5, p. 84).

Cette tendance est probablement également liée à la constitution même du corpus qui est composé de nombreux textes de géographie : les textes de géographie contiennent 27,7 % de phrases obsolescentes alors que les entrées d'histoire n'en contiennent que 8,1 % (*cf.* tableau 2.10, p. 55).

Les règles décrites confirment nos intuitions et les résultats des statistiques descriptives et plus précisément de l'ACP sur la prise en considération d'indices de types différents, de granularité variable ainsi que le fait de les envisager en combinaison et non de manière isolée. Avant de proposer une mise en perspective linguistique et discursive des combinaisons d'indices mises en lumière par l'ACP et l'apprentissage automatique, nous présentons une évaluation quantitative des résultats de l'outil d'apprentissage automatique.

7.3.3 Évaluation quantitative

Nous avons procédé à triple évaluation des résultats du classifieur :
- tout d'abord, une évaluation sur l'ensemble des données qui compare trois algorithmes de classification différents (*cf.* les classifieurs présentés à la section 7.3.1, p. 198) ;
- puis une comparaison entre différents formats extraits du corpus d'apprentissage (*cf.* la description des vues que nous avons créées sur nos corpus, section 7.3.1, p. 200) ;
- enfin, une évaluation du système par les experts du domaine est menée : application des règles sur un corpus de test (*cf.* chapitre 9, p. 217).

Évaluation sur le corpus complet : comparaison d'algorithmes

Les résultats sont évalués selon la méthode de la ***validation croisée***. Cela consiste à diviser les données en k $partitions$, à apprendre sur $k-1$ $partitions$, puis à utiliser la partition écartée, appelée *ensemble test*, pour la phase de test. Le processus est réitéré en écartant une partition différente à chaque fois. Pour nos expériences, nous avons pris $k = 10$, ce qui constitue un paramétrage standard.

L'évaluation de la performance d'un classifieur automatique se fait généralement à l'aide du **score de classification**, c'est-à-dire la proportion d'objets bien classés. Pourtant, cet indice a peu de sens lorsque les classes ne sont pas équilibrées, comme c'est le cas pour notre problème.

Pour chaque classe, les mesures de **rappel** (soit la proportion d'exemples découverts), de **précision** (soit la proportion d'exemples correctement attribués) et leur moyenne harmonique, le **F-score** sont des scores traditionnellement utilisés dans les systèmes de recherche d'information (*cf.* section 1.2, p. 17). Dans notre cas, pour la classe obsolescente, le rappel est de 78 % et la précision de 34 % : ce résultat est encourageant, d'autant plus que dans un contexte *encyclopédique* et parce que le choix final reste au rédacteur, il vaut mieux favoriser le rappel que la précision. En effet, le fait d'oublier une révision est plus grave que d'indiquer inutilement un paragraphe à réviser. Compte tenu de ces performances et des proportions de classe, notre méthode permet de diminuer de deux tiers la tâche du correcteur humain.

Cependant, le rappel et la précision sont caractéristiques d'une classe et non de l'ensemble. Ces mesures évaluent la performance *statique* (*i.e.* pour un seuil fixe) du classifieur, qui peut être variablement interprétée. En faisant évoluer la confiance accordée au classifieur, il est possible de construire une **courbe dite de ROC** Fawcett (2003). Elle représente pour chaque seuil de confiance le taux de *vrais positifs* (proportion d'éléments bien classées pour la classe positive) et de *faux positifs* (proportion d'éléments mal classées). Cette courbe permet d'optimiser, au choix, le rappel ou la précision, selon les besoins de l'application en faisant varier un seuil de 0 à 1 : à chaque seuil sont calculés les taux de vrais positifs et de faux positifs[29].

En d'autres termes, la courbe ROC permet de visualiser le pouvoir séparateur d'un modèle en représentant le pourcentage d'événements bien détectés (les vrais positifs, sur l'axe Y) en fonction du pourcentage d'événements détectés à tort (les faux positifs, sur l'axe X) lorsque l'on fait varier le seuil de séparation du score. Si cette courbe coïncide avec la diagonale, alors le modèle n'est pas plus performant qu'une notation aléatoire. Plus cette courbe est proche du coin supérieur gauche, meilleur est le modèle. On peut superposer plusieurs courbes ROC pour montrer l'apport progressif de chaque variable explicative dans un modèle comme nous le faisons dans la figure 7.25 (p. 207).

C'est également une mesure qui est indépendante de la répartition des classes et permet donc de palier au déséquilibre des classe $obsol\ vs\ non-obsol$. La mesure de l'aire qu'elle définit est un indicateur fiable de la performance du classifieur.

Comme nous l'avons indiqué à la section 7.3.1 (p. 198), nous avons expérimenté trois types de règles pour construire un classifieur à partir de nos indices. L'aire sous la courbe ROC obtenue pour chacun des types de règles est indiqué dans le tableau 7.3.

Ces résultats (autour de 79% quand l'aléatoire donne 50%) montrent que les indices linguistiques exploités sont pertinents pour la caractérisation de l'obsolescence. On constate que les écarts de performance entre les trois classifieurs sont marginaux et que les règles classiques obtiennent les meilleures performances.

[29] Nous étalonnons le classifieur à 0,0501 ce qui permet de privilégier le rappel par rapport à la précision).

	Aire sous la courbe ROC
règles d'association classique	79,8 %
règles généralisées	79,2 %
motifs émergents	76,5 %

TAB. 7.3 - *Comparaison de trois types de règles d'association*

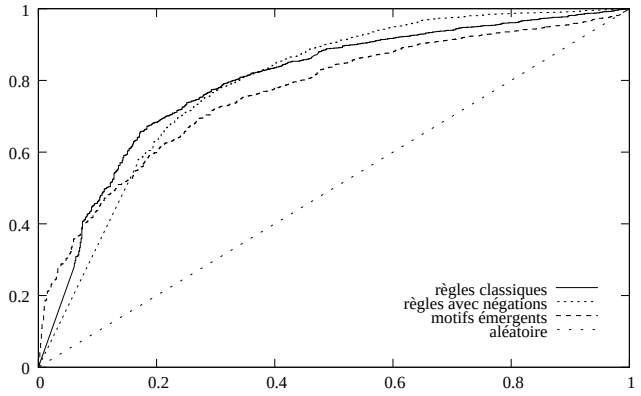


FIG. 7.25 - *Courbes ROC des différents classifieurs. En abscisse le taux de faux positifs, en ordonnée le taux de vrais positifs. Chaque point est obtenu en seuillant différemment la probabilité indiquée par le classifieur.*

C'est d'ailleurs sur elles que nous nous sommes basée pour décrire les combinaisons d'indices dans la section précédente.

La figure 7.25 (p. 207) représente les trois courbes correspondantes. La diagonale représente la performance d'un classifieur aléatoire.

Évaluation de l'apport des différents types d'indices : comparaison de vues sur le corpus d'apprentissage

L'objectif est ici de mesurer l'impact des différents indices et niveaux d'indices pour le repérage automatique de l'obsolescence.

Pour cela, nous avons créé cinq vues différentes du corpus [ENCYCLO] (*cf.* section 7.3.1, p. 200) :

- *corpusComplet* : une vue qui prend en compte tous les indices ;
- *corpusIPseuls* : une vue qui prend en compte uniquement les indices intra-phrastiques ;
- *corpusIPHierar* : une vue qui prend en compte les indices intra-phrastiques et les indices hiérarchiques ;
- *corpusIPPos* : une vue qui prend en compte les indices intra-phrastiques et les indices positionnels ;

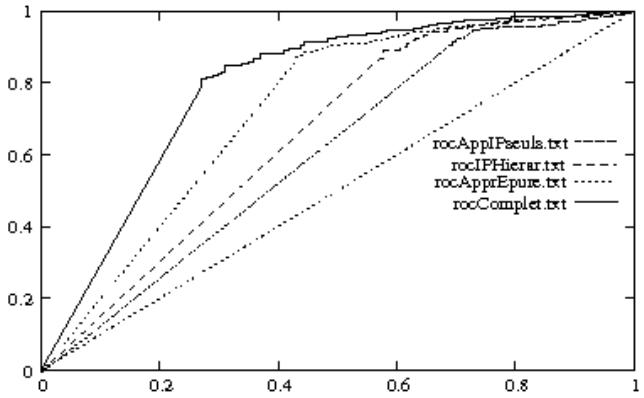


FIG. 7.26 - *Courbes ROC des différentes vues sur le corpus. En abscisse le taux de faux positifs, en ordonnée le taux de vrais positifs. Chaque point est obtenu en seuillant différemment la probabilité indiquée par le classifieur.*

– *corpusEpure* : un corpus « épuré » dans lequel sont enlevées les variables non significatives (en fonction des résultats des statistiques de base, DESCO, et de l'ACP).

L'algorithme utilisé pour l'analyse de ces cinq corpus d'apprentissage est celui des **règles d'association classiques** précédemment utilisé car il est le plus performant (*cf.* tableau 7.3)[30].

Le prototype a été paramétré de la même manière pour les cinq vues d'apprentissage comme nous l'avons présenté à la section 7.3.1 (p. 200).

Les résultats des performances de l'algorithme en précision/rappel sur les différentes vues sont présentées dans le tableau 7.4.

	Précision	Rappel	F-Score	Courbe ROC
corpusComplet	32.9	**78.8**	46.4	**79.23**
corpusIPseuls	38	37	37.5	61.5
corpusIPHierar	39.9	45.6	42.5	66.16
corpusIPPos	33.2	56.7	41.9	68.63
corpusEpure	**38.7**	**62.3**	47.7	**72.96**

TAB. 7.4 - *Comparaison des performances du classifieur (mvminer) selon les différentes vues sur le corpus d'apprentissage*

Les résultats des performances de l'algorithme sur les différentes vues selon la courbe ROC sont présentés dans le tableau 7.4.

Que ce soit selon les mesures de précision/rappel ou à la lecture des courbes ROC, Les meilleurs résultats globaux sont obtenus lorsque tous les types d'indices,

[30]Plus spécifiquement, il s'agit du prototype MVMINER[31].

linguistiques et discursifs, sont utilisés pour l'apprentissage des règles.

D'une manière générale, tous les taux de précision sont bas. Il est vrai que, relativement à l'objectif final, nous avons privilégié un paramétrage de l'algorithme favorisant le taux de rappel. Nous pensons en effet qu'il vaut mieux un outil qui repère le plus possible de phrases obsolescentes même s'il produit des résultats bruités, qu'un outil qui en repère très peu même si elle sont précises[32].

Prendre en compte les résultats des statistiques descriptives (DESCO) pour créer un corpus où seuls les indices significativement présents dans les segments d'obsolescence (*corpusEpure*) sont utilisés est un choix qui produit de meilleurs résultats que celui qui consiste à trier les indices selon leur type (intra-phrastiques, hiérarchiques ou positionnels).

Les résultats les moins bons sont obtenus lorsque seuls les indices intra-phrastiques sont pris en compte. L'apport des indices hiérarchiques et positionnels est faible : la prise en compte des indices hiérarchiques favorisent la précision alors que les indices positionnels privilégient le rappel.

Pour conclure, l'algorithme d'apprentissage automatique que nous avons utilisé est entièrement paramétrable selon les objectifs recherchés (on peut étalonner le classifieur comme on le souhaite). Pour notre part, nous avons fait le choix de privilégier le rappel mais nous pourrions également chercher à favoriser le taux de précision.

[32] Un outil industrialisé laisserait dans tous les cas l'ultime choix décisionnel au rédacteur humain. Dans ses conditions, il est donc préférable de favoriser le rappel.

Chapitre 8

Discussion sur les combinaisons d'indices repérées dans les segments d'obsolescence

Ce chapitre récapitule les combinaisons d'indices qui ont émergées de l'Analyse en Composantes Principales et de l'Apprentissage Automatique. Nous proposons de relier nos observations spécifiques sur les segments d'obsolescence avec des travaux en T.A.L. traitant également de la question de la multiplicité des indices.

D'une manière générale, les conclusions issues de l'Apprentissage Automatique sous forme de règles d'association (*cf.* section 7.3.2) confirment les tendances observées dans l'analyse des résultats des statistiques descriptives (*cf.* section 7.1.2) et de l'ACP (*cf.* section 7.2.2).

Ces tendances sont perceptibles sur plusieurs points :

- la question de validité/pertinence des types d'indices utilisés pour le repérage de l'obsolescence (en termes de caractérisation sémantique : temps, point de vue du locuteur, types de référent dans les entités nommées, aspect et modalité principalement) ;
- la question de la position des indices dans la phrase ;
- la prise en compte d'indices positionnels (structure du documents, indices à gros grain) ;
- les indices hiérarchiques (*i.e.* les titres) ;
- les indices externes (*i.e.* les rubriques) ;

Les remarques de ce chapitre tendent à répondre à deux objectifs : d'une part évaluer la pertinence des indices en eux-mêmes mais aussi et surtout valider les combinaisons d'indices qui ont émergé des analyses statistiques et ainsi fournir une description de l'obsolescence et des segments d'obsolescence plus précise.

8.1 Le typage sémantique des indices

Le typage sémantique des indices concerne la nature et l'annotation sémantique des expressions linguistiques repérées. Avant de procéder à l'interprétation des résultats statistiques, nous avions une intuition forte sur l'importance du temps, des aspects modaux et aspectuels, de la présence d'entités nommées ou encore de l'expression du point de vue.

En ce qui concerne les indices temporels, on observe deux tendances diamétralement opposées. La première, orientée du côté de l'obsolescence, insiste sur l'importance des indices temporels de type *déictique* et/ou référant au moment d'énonciation du rédacteur (*coïncidence* et *postériorité*). La tendance inverse, montre que des informations temporelles de type *antériorité* sont corrélées négativement à l'obsolescence (*cf.* les statistiques descriptives).

Ces indices sont des indices forts même pris isolément. Envisagées en configuration avec d'autres indices, ils rendent la fiabilité du jugement d'obsolescence plus forte.

Concernant les entités nommées, la tendance est identique. Il existe une sémantique de base propre aux segments d'obsolescence. C'est notamment le cas lorsqu'une entité nommée de type *géopolitique* ou de type *mesure* est présente dans un segment textuel. Par contre, la présence d'une entité de type *personne* aura tendance à influencer l'interprétation sur la classe non-obsolescente (s'il n'y a pas d'autre(s) indice(s) fort(s) orientant vers l'obsolescence).

Le cas des entités nommées de *lieu ville* ou/et *pays* est plus mitigé : la présence d'autres éléments dans le contexte phrastique est nécessaire pour juger de l'obsolescence du segment. Ainsi, l'obsolescence sera plus probable si l'entité nommée de *lieu* est associée à une entité nommée de type *géopolitique* et/ou de type *mesure* ou une expression temporelle de *coïncidence* ou *postériorité*.

Concernant les temps verbaux, le conditionnel est corrélé positivement à l'obsolescence (statistiques descriptives, ACP et AA). Les règles d'association mettent en avant l'usage du passé composé lorsqu'il est associé à une expression temporelle de type *déictique coïncidence* (*cf.* section 3.2, p. 68).

Quant au futur, la corrélation est quasiment nulle avec l'obsolescence[1] (*cf.* section 3.2.1, p. 69). Ainsi, l'association intra-phrastique avec le conditionnel par exemple ne ressort pas comme un bon marqueur de l'obsolescence (il faudrait pouvoir rechercher les associations interphrastiques (*cf.* section 6.2, p. 167).

Les valeurs aspectuelles n'apparaissent pas corrélées de manière fiable à l'obsolescence, que ce soit négativement ou positivement (tant dans l'ACP que dans

[1] La variable n'apparaît que dans le facteur 8 de l'ACP ; le descripteur est très peu présent dans les règles d'association.

l'AA).

On peut alors se demander s'il est nécessaire de conserver ce type d'indice qui n'apparaît pas dans les combinaisons d'indices. Au contraire ne faudrait-il pas affiner la catégorie pour la rendre plus pertinente et plus adaptée au but recherché. Cette question de la conservation et affinage des indices ou de leur suppression mérite d'être posée si ce projet se poursuit.

Les informations concernant le type de point de vue du locuteur sur ses propos est important, dans deux cas principalement : lorsque l'auteur fait une prévision sur un événement et lorsqu'il se distancie de ses propos en citant ses sources. Le premier cas s'explique à travers la question du rapport au temps, de l'introduction d'un monde hypothétique. Dans le second cas, lorsqu'une source est signalée, elle l'est souvent pour justifier la validité d'une information (souvent une valeur chiffrée) de la phrase. La catégorisation en termes de point de vue du locuteur mériterait d'être reprise et approfondie, à la fois en termes de types de catégories mais également dans le choix des expressions langagières appartenant à ces classes.

Concernant les argumentatifs, alors que l'ACP ne révèle aucune relation (fiable) avec l'obsolescence, les règles d'association montrent l'intérêt des connecteurs exprimant une identité (« comme »), une explication (« ainsi », « car »), une opposition (« au contraire ») ou encore une correction (« mais »).

Marcu (2000) ou Hernandez (2004) constatent que les connecteurs discursifs sont des indices fiables pour la segmentation thématique des textes alors que les indices lexicaux sont plutôt peu discriminants.

Dans notre cas, les indices lexicaux, parce qu'ils apportent un point de vue sémantique particulier orienté vers la question de l'obsolescence, sont des indices centraux pour le repérage des segments obsolescents alors que les connecteurs discursifs (*i.e.* les argumentatifs dans notre travail) ne représente qu'un aspect parmi d'autres.

Nos conclusions vont ainsi plutôt dans le sens des travaux de Bouffier (2008) : l'auteur insiste sur l'importance de la relation entre des indices lexicaux et des indices visuels.

8.2 Les indices positionnels phrastiques

Sur la position des indices au sein de la phrase, l'ACP et l'AA mettent en avant l'importance des positions normale[2] et finale. Contrairement à ce que nous attendions, le fait qu'un indice soit en initiale de phrase ne préfigure pas le caractère obsolescent du segment dans lequel il apparaît.

[2]Nous parlons de position normale lorsqu'un indice n'est ni en initiale ni en finale. C'est la position par défaut de chaque indice.

Ces résultats vont à l'encontre des hypothèses formulées dans la section 2.4.2 concernant le potentiel structurant des indices à l'initiale de la phrase : nous supposions alors qu'un indice à l'initiale de la phrase pouvait contribuer à l'identification des segments obsolescents. Cette hypothèse était notamment liée à la notion de cadre d'interprétation et à leur repérage.

Sur ce point, il faut rester prudent quant aux conclusions à tirer. En effet, nous avons souligné dans la section 6.2 notre incapacité à rendre compte du phénomène de la portée sémantique des éléments en position initiale de phrase. Au stade de notre travail, nous travaillons sur l'unité *phrase*. Or le phénomène de portée se définit par son caractère supra-phrastique. Nous avons proposé un moyen de rendre compte de ce phénomène dans cette même section 6.2 (p. 167).

8.3 Les indices positionnels textuels

Les indices positionnels textuels rendent compte de la position de la phrase dans le paragraphe et du paragraphe dans le document.

Dans l'ACP, la **position de la phrase dans le paragraphe** n'est pas une information forte. Dans les règles d'association en revanche, les première et dernière phrases de paragraphes sont mises en avant :

- les premières phrases de paragraphe associées à des indices temporels de type *coïncidence* ou à des entités nommées de type *mesure*.
- les dernières phrases de paragraphe associées au conditionnel et à des entités nommées de type *géopolitique*.

Il s'agit de mettre en évidence des stratégies de mises en texte spécifiques : nous supposons un lien entre la présence de segments obsolescents et la structuration propre du texte, son organisation logique. Si traditionnellement on distingue le positionnement textuel (*i.e.* la structure logique du document) de la structure rhétorico-sémantique du document, dans notre cas, nous posons le présupposé[3] suivant lequel, par exemple, le premier paragraphe d'une section est une introduction et que le dernier paragraphe d'une section est une conclusion.

HoDac (2007) évoque l'importance des éléments en première phrase de paragraphe ou de section sur la variation textuelle.

Sur la **position des paragraphes dans le document**, la position d'introduction de section est souvent corrélée à des indices temporels ou des entités nommées de type *géopolitique* alors que la position conclusive est associée aux entités nommées de type *mesure*. Ces observations montrent qu'un paragraphe introductif va avoir tendance à situer le décor temporel ou référentiel (*i.e.* de quoi on va parler).

Un paragraphe conclusif a un rôle différent : on conclut souvent sur la base d'exemples précis, concrets qui font appel à des valeurs chiffrées précises.

[3] Ce parti-pris est fort et peut être discuté et critiqué. Il correspond selon à une réalité éditoriale précise au sein de laquelle les procédés de mise en page sont relativement stricts et définis non pas par les rédacteur eux-mêmes mais mais la politique éditoriale globale.

Ce constat va dans le sens des travaux de Marcu (2000) notamment qui a travaillé sur la relation entre la présence de marques linguistiques particulières et leur apparition dans des positions paragraphiques précises pour juger automatiquement de l'*importance* d'une phrase (pour un système de résumé automatique).

Bouffier (2008) met également en avant le fait que les indices relevant de la mise en forme matérielle (position dans le paragraphe et position dans le document) sont des éléments discriminants pour son objectif (recherche des relations conditionnelles entre segments, *cf.* section 1.3.3)

8.4 Les indices hiérarchiques

Les indices hiérarchiques, *i.e.* les indices présents dans un titre, sont des indices pertinents pour la recherche de segments obsolescents. L'ACP et l'AA convergent vers le même constat : une entité nommée de type *géopolitique, mesure ou lieu* ou une expression temporelle de type *déictique* dans un titre est significativement corrélée à l'obsolescence.

De la même manière que Ibekwe-SanJuan (2005), nous constatons une faible présence d'indices de rhétorique (*i.e.* de connecteurs discursifs) et d'indice de nouveauté (*i.e.* de point de vue) dans les titres.

8.5 Les indices externes

La caractérisation du texte en fonction des rubriques thématiques est enfin une information centrale : que ce soit à travers les statistiques descriptives ou l'AA, les résultats montrent que certains indices ou combinaisons d'indices sont pertinents pour une rubrique particulière. L'opposition la plus marquée concerne les textes relevant de la rubrique histoire et ceux relevant de la rubrique géographie : alors qu'une combinaison d'indices comme « *entité nommée lieu* + *entité nommée mesure* » sera fortement associée à l'obsolescence dans un texte géographique, la même combinaison sera contre-productive en histoire.

Ce constat va dans le sens des travaux de Zerida *et al.* (2006) même s'il s'agit plutôt d'une classification des textes en types : les auteurs constatent une différence significative dans l'organisation de l'écrit et dans le style de trois types de textes biomédicaux (articles de recherche, de synthèse, cliniques, *cf.* section 1.3.3).

8.6 Conclusion

La description des segments d'obsolescence est mieux délimitée. La méthodologie développée nous a permis de porter un regard objectif sur le phénomène de l'obsolescence à travers, d'un côté, les types d'indices utiles au repérage de l'obsolescence, et de l'autre, les combinaisons pertinentes dans les segments obsolescents. D'une manière générale, l'hypothèse selon laquelle l'obsolescence est repérable

par des indices linguistiques est validée ainsi que la nécessité de les envisager en combinaisons.

Notre ultime objectif consiste à évaluer la pertinence de nos observations et résultats en proposant à des experts de juger d'une annotation automatique des segments d'obsolescence dans un corpus nouveau. C'est ce qui est présenté dans le chapitre 9 suivant.

Chapitre 9

Évaluation par les experts

Nous avons finalement procédé à une évaluation du système à travers le regard des experts, utilisateurs finaux d'un outil d'aide à la mise à jour de documents.

À partir des règles apprises automatiquement par le classifieur, des textes « nus » ont été annotés automatiquement : pour chaque phrase du corpus de test est indiqué son caractère obsolescent ou non. Nous avons préalablement annoté ce corpus de test à l'aide de l'outil ALIDIS (*cf.* présentation générale de la méthode menée à la page 2). Le format des données des corpus de test est présenté dans l'annexe E (p. 301).

9.1 Résultats en termes de performance du classifieur

Nous considérons qu'une phrase est obsolescente lorsque le seuil de score associé est supérieur à 0,0501[1] comme nous l'avons présenté dans la section 7.3.3 (p. 206). Il s'agit d'un étalonnage du classifieur qui nous permet de choisir, sur la base d'un seuil donné, de favoriser le rappel par rapport à la précision.

Nous avons comptabilisé les intersections des annotations automatiques avec la validation humaine des 3810 phrases du corpus de test. Les résultats sont fournis dans le tableau 9.1 :

		humain		
		obsol	non obsol	total
machine	obsol	261	440	701
	non obsol	138	2971	3109
	total	399	3411	3810

TAB. 9.1 - *Intersections des annotations automatiques avec la validation humaine*

Ces chiffres nous indiquent que, toutes classes confondues, 85 % des cas, qu'ils soient obsolescents ou non, sont traités correctement par la machine. Si on applique

[1] Ce qui implique une proportion de classe de 0,1960 : cela signifie que le classifieur renvoit presque 20 % de phrases obsolescentes.

le cœfficient Kappa, on observe un taux d'accord entre l'humain et la machine de 0.39, score qui reste assez proche et cohérent des scores d'accord entre les juges (situés entre 0.35 et 0.50, *cf.* section 2.4.1, p. 50).

En termes de scores de pertinence, le tableau 9.2 rend compte de l'évaluation du classifieur automatique pour la classe *obsol*.

	Précision	Rappel	F-Score
corpusComplet	0,37	0,65	0,47

TAB. 9.2 - *Évaluation par les experts : les résultats complets*

En d'autres termes, l'outil identifie correctement deux phrases obsolescentes sur trois et considère comme obsolescentes deux phrases sur dix. Ce sont des résultats plutôt encourageants dans un contexte professionnel : on observe une réduction du nombre de phrases à analyser même si un certain nombre nous échappent. Nous proposons dans la section suivante d'évaluer et décrire avec précision quelles phrases ne sont pas annotées obsolescentes : est-ce parce qu'un indice est mal repéré, mal annoté ?

Les observations des experts nous amènent également à évaluer l'outil en fonction du type de texte (*i.e.* selon la rubrique). Les résultats de cette évaluation sont fournis dans le tableau 9.3.

	Précision	Rappel	F-Score
Sport	0,33	0,33	0,33
Économie-Politique	0,65	0,52	**0,58**
Médecine	0,30	0,43	0,35
Histoire	0,01	0,13	0,03
Société	0,10	0,33	0,15
Géographie	0,69	0,84	**0,76**
Faune et Flore	0,07	0,86	0,13
Arts et Littératures	0,45	0,65	**0,54**

TAB. 9.3 - *Évaluation par les experts : les résultats par rubrique*

Ces chiffres montrent que l'outil est plutôt bon pour ce qui est du repérage de l'obsolescence dans les fiches traitant de géographie (les textes de géographie représentent un tiers des phrases du corpus d'apprentissage. En économie-politique et arts et littératures, il est moyen mais les résultats restent corrects. En revanche, il est très mauvais pour les textes d'Histoire, de faits de Société et de Faune et Flore. Il s'agit avant tout d'une limite dûe au corpus d'apprentissage qui mériterait d'être plus volumineux et d'être plus représentatif des segments obsolescents des entrées d'histoire ou de médecine par exemple (pour rappel, dans le corpus d'apprentissage, 27,7 % des segments obsolescents sont de type géographique, contre 8,1 % pour les entrées d'histoire et 7,1 % pour les entrées de médecine (*cf.* tableau 2.10, p. 55).

En d'autres termes, lorsqu'il y a peu de segments obsolescents par nature dans des textes, il n'arrive pas à les repérer efficacement. Un algorithme qui privilégierait la précision sur ce type de textes produirait peut-être de meilleurs résultats.

9.2 Description par l'exemple des erreurs du classifieur

Nous avons relevé un certain nombre d'exemples qui illustrent des cas d'annotation erronée de l'obsolescence et des cas où le segment est obsolescent mais n'a pas été repéré automatiquement. Ces exemples peuvent permettre de mettre au point des améliorations du système de repérage automatique de zones obsolescentes.

9.2.1 Des exemples de segments annotés *obsol* alors qu'il ne le sont pas.

Certaines phrases sont considérées comme obsolescentes suite à un repérage et une annotation sémantique fausse d'un indice par l'outil ALIDIS. Ainsi, dans l'exemple 9.1, c'est à cause de l'erreur d'annotation de l'indice temporel que la phrase est considérée comme obsolescente. En effet, l'expression « avant notre ère » seule est délimitée et annotée comme une expression temporelle de type *déictique coïncidence* (sur la base du possessif « notre »).

Le gnomon : Le plus primitif des cadrans solaires date du IIe millénaire **avant notre ère**. [...]

Source : Corpus Atlas

Exemple 9.1 - *Erreur d'annotation de l'outil* ALIDIS

Dans d'autres cas, ce sont les indices eux-mêmes qui mériteraient d'être annotés différemment. Dans l'exemple 9.2, l'adverbe « désormais » conditionne l'interprétation obsolescente de la phrase : l'étiquette sémantique qui lui est associée est *déictique coïncidence*. Dans ce cas, cet adverbe n'est évidemment pas déictique. Il serait nécessaire, pour ces types de cas-là, soit de reprendre les annotations sémantiques des adverbes dans les lexiques et les grammaires de l'outil ALIDIS, soit de trouver une solution de désambiguïsation de ce type d'adverbe selon leur contexte d'apparition. Par exemple, lorsqu'il y a des dates dans les phrases précédentes, il faudrait propager leur sémantique temporelle sur l'ensemble des phrases de la section.

Parfois, il est difficile d'envisager des solutions vraiment opératoires. Ainsi, dans l'exemple 9.3, deux types de solutions peuvent être mis en place :

- la propagation de la sémantique issue d'expressions temporelles précédentes comme nous l'avons suggéré pour l'exemple 9.2 ;
- introduire de la connaissance encyclopédique sur certaines expressions. Dans cet exemple 9.3, il serait possible de déterminer différentes natures séman-

Les faillites bancaires se multiplient : 642 en 1929, 1 345 l'année suivante et 2 298 en 1931.
Le crédit est dès lors impossible puisque les banques ne répondent plus à la demande.
La fin du crédit génère alors une crise à la fois sociale et économique, ne laissant aucun secteur intact.
Les entreprises industrielles, désormais privées de prêts et de débouchés, doivent limiter leur production, voire fermer.
[. . .]

Source : Corpus Atlas

Exemple 9.2 - *Annotation sémantique des indices à préciser*

tiques possibles pour le superlatif ou créer un lexique de noms propres contenant des expressions comme « campagne de Saxe ». Cette solution orientée connaissances non linguistiques semble difficile à mettre en œuvre.

La campagne la plus longue : La campagne de Saxe, 28 ans. [. . .]

Source : Corpus GLI

Exemple 9.3 - *Introduction de connaissances non linguistiques*

Enfin, pour un certain nombre de phrases il est difficile de dire si l'information est obsolescente ou non. C'est le cas typique où l'information potentiellement obsolescente est discutable et où il serait souhaitable de considérer l'obsolescence comme un phénomène graduable. Ainsi, l'exemple 9.4 a peu de risque de devoir être mis à jour mais pourtant c'est une information susceptible d'évolution. Le cas est identique dans l'exemple 9.5.

L'Asie centrale est composée du Kazakhstan, du Kirghizistan, de l'Ouzbékistan, du Tadjikistan, du Turkménistan (cinq républiques issues de l'Union soviétique et membres de la Communauté des États indépendants, ou CEI, fondée en 1991) et de la Mongolie. [. . .]

Source : Corpus GLI

Exemple 9.4 - *Le caractère obsolescent de l'information est discutable - 1*

Entre les forêts de l'Argonne et le versant ouest du massif des Vosges, la Lorraine offre des paysages verdoyants : 36 % des terres sont recouverts par des forêts, et l'eau y abonde. [. . .]

Source : Corpus GLI

Exemple 9.5 - *Le caractère obsolescent de l'information est discutable - 2*

9.2.2 Des segments *obsol* qui n'ont pas été repérés

Les cas où une phrase contenant de l'information obsolescente n'est pas repérée concernent un cas sur trois. Comme pour les segments mal repérés, il est possible, au travers d'exemples ciblés, d'envisager des améliorations du système.

Les expressions temporelles devraient pouvoir jouer un rôle plus important. Dans l'exemple 9.6, l'adverbe « aujourd'hui » devrait pouvoir être fortement pondéré afin de conduire l'outil vers une interprétation de l'obsolescence.

> Il est encore populaire aujourd'hui, notamment en Inde.. [. . .]
>
> Source : Corpus GLI

Exemple 9.6 - *Pondérer fortement certains indices (temporels par exemple)*

Comme pour les repérages erronés des segments d'obsolescence, la possibilité de propagation de traits temporels permettrait sans doute de considérer une phrase comme celle de l'exemple 9.7 comme obsolescente.

> Mais la culture arabe ne se réduit pas à son aspect religieux, et, si l'islam compte plus d'un milliard d'adeptes à travers le monde, seul un cinquième d'entre eux sont de langue arabe. [. . .]
>
> Source : Corpus GLI

Exemple 9.7 - *Propagation des traits temporels présents dans une phrase précédente*

Enfin, des règles contextuelles du niveau de la phrase et de la section pourraient améliorer certains cas. En effet, au milieu de segments obsolescents, il faudrait pouvoir considérer les autres segments comme obsolescents.

Dans l'exemple 9.8, les phrases en gras ne sont pas annotées obsolescentes alors qu'elles le sont. Ici encore, la propagation des traits temporels permettrait sans doute d'améliorer les résultats du système.

> **Des campagnes dynamiques**
> La richesse des sols est très favorable à une agriculture variée même si celle - ci a reculé comme ailleurs en termes d'emploi et de PIB.
> On cultive en Alsace du maïs, des céréales, du chou à choucroute et de la betterave à sucre.
> **Outre le houblon, la région produit aussi du tabac.**
> L'horticulture et les vergers alsaciens sont célèbres.
> **La surface cultivable, estimée à 300 000 ha, est divisée en propriétés de petite taille.**
> L'élevage assure une production de viande et de lait non négligeable.
> La vitrine de l'Alsace reste indéniablement la viticulture.
> La qualité des sols associée à un bon ensoleillement offre à la vigne des conditions idéales.
> [. . .]
>
> Source : Corpus GLI

Exemple 9.8 - *Proposer des règles contextuelles larges*

Bilan de la troisième partie

Dans cette troisième (et dernière) partie, nous avons décrit la méthodologie mise en œuvre pour décrire les segments d'obsolescence et automatiser le repérage de segments d'obsolescence dans les textes.

La méthodologie RIO (pour Repérage d'Informations Obsolescentes) est composée de trois outils principaux :

- l'outil ALIDIS a pour objectif le repérage et l'annotation sémantique d'expressions linguistiques spécifiques (*cf.* chapitre 5) ;
- l'outil OCAS permet la transformation du format des données en sortie de l'outil ALIDIS pour un traitement descriptif et statistique (*cf.* chapitre 6) ;
- l'outil STAAT produit des analyses statistiques variées sur les données issues de l'outil OCAS (*cf.* chapitre 7).

L'outil STAAT rassemble des procédures, logiciels et prototypes externes existant indépendamment de notre travail[2]. Les statistiques descriptives (*cf.* section 7.1), l'ACP (*cf.* section 7.2) et la fouille de données (*cf.* section 7.3) permettent, chacun à leur niveau, de mettre en évidence des fonctionnements textuels et linguistiques intéressants des indices dans les segments d'obsolescence.

Ainsi, le cœfficient de corrélation explore les rapports entre la variable *obsol* et toutes les autres variables possibles existant dans la base de données d'apprentissage. Ces calculs statistiques valident la pertinence des indices repérés par ALIDIS dans les segments obsolescents. Tous les types et niveaux d'indices sont d'ailleurs représentés. Ces calculs statistiques permettent également un tri éventuel sur ces mêmes indices. C'est notamment sur la base de ces résultats que la vue *corpusEpure* a été créée (*cf.* section 7.3.1).

L'Analyse en Composantes Principales a mis en évidence un certain nombre de corrélations entre des indices et la variable *obsol*[3]. Cette analyse de données, dont les résultats et interprétations doivent être considérées avec prudence au vu des faibles valeurs, nous a permis d'approfondir la réflexion sur les combinaisons d'indices possibles dans les segments d'obsolescence. L'analyse des axes 2 à 7 nous a conduite à créer quelques règles de décision : les résultats sont faibles en termes de précision et de rappel (de de F-score également) et l'interprétation des

[2]Nous avons notamment utilisé le logiciel SPAD et le prototype MVMINER.

[3]Nous avons fait le choix de faire l'ACP sur l'ensemble des données que la variable *obsol* soit vraie ou fausse. Il serait intéressant de produire deux nouvelles ACP, une qui ne prendrait en compte que les segments obsolescents, l'autre qui ne considérerait que les segments non obsolescents.

axes et leur transformation en règles s'est avérée une étape fastidieuse. Un travail complémentaire consisterait à produire une classification ascendante hiérarchique sur la base des résultats de l'ACP (le logiciel SPAD offre cette possibilité).

Dans cette optique et parce que nous préferions un modèle exploratoire, nous avons finalement utilisé l'outil d'apprentissage automatique MVMINER développé par François Rioult. L'utilisation de ce type d'outil est doublement motivée : d'une part la description des segments d'obsolescence et celle des combinaisons d'indices a été approfondie ; d'autre part, les règles apprises par le prototype ont été appliquées sur un corpus de test ce qui a permis d'évaluer les performances d'un prototype d'aide à la mise à jour de textes encyclopédiques.

Une discussion sur les connaissances extraites concernant les combinaisons d'indices a permis de mettre en relief l'importance de considérer des indices linguistiques et discursifs. Les observations ont également été mises en perspectives par rapport à des travaux connexes en T.A.L.

Une double évaluation du prototype d'aide à la mise à jour a finalement été mise en place :

- une évaluation intrinsèque du système par validation croisée qui a produit des résultats plutôt encourageants ;
- une évaluation par les experts qui a confirmé cette tendance.

Ces évaluations fournissent une estimation *a posteriori* sur la qualité de l'outil ALIDIS et sur les possibilités de l'améliorer : c'est sur l'outil ALIDIS que tout repose car c'est lui qui crée les indices pertinents pour le repérage de l'obsolescence. Ainsi, la plupart des améliorations consisteront à modifier et adapter le repérage des indices linguistiques, sémantiques et structurels.

Dans l'ensemble, tous ces outils ont rendu possible une description précise des segments d'obsolescence, des indices susceptibles de les délimiter et des combinaisons d'indices pertinentes à l'intérieur de ce type de segments.

Nous avons réalisé une chaîne de traitement complètement automatique et fonctionnelle qui identifie les segments obsolescents. L'ensemble des traitements est également entièrement reproductible.

Conclusions et Perspectives

Rappel des objectifs de cette thèse

Ce travail vise un objectif appliqué concret qui prend racine dans le domaine de l'édition d'encyclopédies et qui vise la notion de mise à jour de l'information. Pour répondre à ce besoin, nous supposons que l'automatisation de la tâche de mise à jour est possible : nous avons montré que le contexte scientifique et technique dans lequel nous nous inscrivons en permet l'élaboration.

L'obsolescence : une définition plus précise

La tâche de mise à jour est décrite à travers le phénomène d'obsolescence : un segment textuel est susceptible de devoir être vérifié et mis à jour s'il contient une information obsolescente, c'est-à-dire qui est susceptible d'évoluer dans le temps, d'être devenue fausse.

Ainsi nous avons proposé la distinction entre les réactualisations et les réadaptations. Cette opposition n'a malheureusement pas été développée comme nous l'espérions.

Idéalement, s'il est convenu avec les rédacteurs que cette distinction est pertinente pour la tâche[4], il serait intéressant de proposer un typage, une caractérisation linguistique précise des segments obsolescents qui correspondent à cette distinction sur la base des nouvelles connaissances que ce travail apporte sur l'obsolescence. Concrètement, cette distinction peut être envisagée comme le moyen de gérer l'urgence des mises à jour dans le prototype final : il nous semble qu'il serait alors plus urgent de vérifier le contenu informationnel des réadaptations car ce sont elles qui sont le plus à même de véhiculer des informations réellement fausses tandis que les réactualisations informent sur des faits qui seront *a priori* toujours vrais mais dont la pertinence temporelle peut être discutée.

Pour donner un exemple concret, il nous semble qu'il est important de vérifier et éventuellement de modifier une phrase comme « Il n'existe pas à l'heure actuelle de vaccin contre le SIDA. » si, depuis la date d'édition, un vaccin a été trouvé ; alors que dans une phrase comme « En 2004, le PIB de la France est de 27 600 dollars. », l'information est vraie et elle le sera toujours mais elle ne sera vraissemblablement plus très pertinente pour un lecteur de 2008.

Nous avons insisté sur le caractère fondamentalement non-linguistique de l'obsolescence pour laquelle il n'existe pas de marqueurs linguistiques dédiés : c'est pourquoi nous avons orienté notre travail vers la recherche de combinaisons d'indices. C'est alors la combinaison en elle-même qui est susceptible d'être un bon (ou mauvais) marqueur de l'obsolescence.

[4] Ce point doit être développé avec les rédacteurs et évalué précisément.

Repérer les segments d'obsolescence : des indices discursifs

La description de l'obsolescence à travers l'étude approfondie des segments d'obsolescence annotés manuellement par des experts a mis en évidence un certain nombre d'indices linguistiques et discursifs récurrents.

Notre approche vise à faire travailler ensemble des indices pertinents multiples : des indices temporels, aspectuels, modalisateurs, des entités nommées ou encore des éléments relevant de la typo-disposition des textes peuvent, ensemble, devenir des marqueurs de l'obsolescence.

Les aspects temporels sont centraux lorsqu'on s'intéresse à l'obsolescence. C'est en effet sur eux que le découpage des événements et la relation à la situation d'énonciation est d'abord perçue. Nous avons également montré que le temps n'est pas le seul aspect à prendre en compte : les valeurs aspectuelles et modaux se sont révélés des pistes de recherche intéressantes. Le point de vue du rédacteur, souvent exprimé implicitement, semble également un critère pertinent pour évaluer la validité d'une information. Enfin, la très large et hétérogène catégorie des entités nommées (valeurs chiffrées, lieux, noms propres, lexiques spécifiques, etc.) regroupe des expressions linguistiques qui semblent productives pour l'obsolescence.

La large variété sémantique des indices est donc un premier aspect fondamental pour aborder le repérage de l'obsolescence.

L'outil ALIDIS a pour objectif le repérage systématique de tels indices dans notre corpus [ENCYCLO]. La performance de cet outil est correcte mais une amélioration des repérages et des annotations est souhaitable.

Nous envisageons deux types d'évolutions. Premièrement, une évolution plutôt d'ordre technique visant à réduire les taux de bruit et de silence est prévue. Deuxièmement, une redéfinition des indices (re-classification, affinage de traits, fusion de types d'indices, suppression ou ajout d'indices) devra être menée sur la base des résultats fournis par les traitements statistiques (ACP et règles d'association).

De manière générale, les résultats de ce travail tendent à confirmer la pertinence d'une approche prenant en considération des indices différents et à granularité variée. Dans la section 7.3.3, nous avons ainsi montré l'intérêt de la prise en compte d'indices hiérarchiques et d'indices positionnels textuels.

Concernant les indices positionnels phrastiques, il est nécessaire d'ajouter un analyseur de cadres qui permettrait de gérer le phénomène de portée sémantique des indices à l'initiale de phrase. Le principe d'héritage du contexte (Zerida *et al.*, 2006) serait alors appliqué comme nous le faisons pour les titres et les indices à gros grain (*i.e.* les indices de type *englobants* dans le modèle 6.2, p. 163).

Le développement et l'exploitation d'un tel analyseur nous permettrait de juger véritablement de l'intérêt d'indices tels que les introducteurs de cadre (IC). Nous supposons que, comme les titres, les IC agissent sur un segment textuel donné et, parce qu'ils véhiculent un sémantisme particulier, ils sont susceptibles de faciliter le repérage de l'obsolescence. De plus, le traitement des cadres de discours nous

semble un moyen intéressant de répondre à l'opposition entre *segment minimal* et *cadre d'interprétation* dont nous avons parlé à la section 2.4.2 (p. 57). Le système actuel ne considère que l'unité phrase alors que la définition d'une unité supérieure se justifie pleinement dans le cas des segments d'obsolescence.

Un travail conséquent a consisté à transformer les données linguistiques et discursives en données traitables statistiquement. Nous avons fait le choix de passer par l'élaboration d'un modèle conceptuel des données (*cf.* p. 163) qui concrètement prend la forme d'une base de données. Ce choix est doublement motivé : *(i)* rendre les traitements reproductibles à d'autres tâches (*i.e.* sur la base d'indices linguistiques différents par exemple) et *(ii)* permettre un accès aisé aux données et la possibilité d'apporter facilement des modifications des données.

Pour notre travail, ce modèle mis en œuvre est suffisant. Il pourrait être enrichi sur la base de la prise en compte des relations non hiérarchiques entre les éléments : nous pensons par exemple au phénomène de l'anaphore qui met en relation des unités de même niveau.

La recherche des combinaisons d'indices est menée dans l'outil STAAT. L'utilisation de méthodes statistiques de fouille de données permet de faire émerger du corpus des informations nouvelles qu'il serait impossible de valider manuellement au vu de la quantité des données dont nous disposons.

Nous avons d'abord mené des statistiques descriptives puis une Analyse en Composantes Principales dans le but d'observer si des variables et/ou des groupes de variables tendaient à être corrélées à la variable obsolescence. Les statistiques descriptives ont montré des tendances intéressantes dans le fonctionnement des variables prises en compte et l'ACP a permis d'aller plus loin en permettant d'envisager les variables non pas de manière binaire mais en termes de configurations, de factorisation des variables.

L'ACP a également prouvé que le phénomène de l'obsolescence est complexe et qu'il ne saurait se réduire à quelques axes forts et donc à quelques indices et/ou combinaisons d'indices uniques. Du fait de cette complexité, nous nous sommes dirigée vers les techniques d'apprentissage automatique.

Notre travail étant exploratoire, nous avons utilisé un système à base de règles d'association. Ce choix nous semble le plus adapté, pour plusieurs raisons. D'abord cette technique est classiquement utilisée dans le cas d'études exploratoires : toutes les associations possibles pour toutes les conclusions de classes possibles sont exploitées, un filtre sur la classe souhaitée étant finalement appliqué (pour nous, les règles concluant sur la classe *obsolescence*. De plus, ce type de système vise l'exhaustivité des associations alors qu'un système comme les arbres de décision vise la recherche d'un arbre idéal. Or notre objectif n'est pas de produire une sélection des indices mais de rechercher les combinaisons, les associations pertinentes d'indices pour l'obsolescence.

De l'intérêt de considérer les indices en combinaisons

Les traitements statistiques et les observations qui en ont découlé ont montré l'intérêt de prendre en compte les indices linguistiques et discursifs en combinaisons. Dans le chapitre 8, nous avons confronté nos observations aux résultats de différents travaux exploitant également des indices linguistiques en combinaison.

Au-delà du fait que certaines intuitions ont été confirmées (entre autres, le temps, certaines entités nommées), que des tendances ont été précisées (la prise en compte des rubriques pour certaines configurations par exemple) et que des relations moins attendues ont été révélées (l'usage du passé composé, l'absence de lien avec le futur, etc.), ces résultats nous donnent les moyens d'approfondir et d'améliorer les indices linguistiques et discursifs à prendre en compte mais aussi de revenir sur la définition du segment d'obsolescence et plus largement de la tâche de mise à jour des textes encyclopédiques.

Perspectives scientifiques et industrielles

Il s'agit d'une étude exploratoire sur le phénomène de l'obsolescence.

Parce que notre objectif vise la recherche de combinaisons d'indices, nous avons fait le choix de couvrir un éventail très large d'indices linguistiques sans nous préoccuper de leur description linguistique précise. Maintenant que la méthodologie mise en place a montré son utilité (envisager les indices linguistiques en complémentarité), il faut maintenant affiner l'ensemble des catégories prises en compte (notamment le temps, l'aspect et la modalité), comparer les diverses théories sur ces questions et mesurer leur apport. Ce travail d'approfondissement peut être mené à deux niveaux : d'abord relativement aux rubriques pour lesquelles il a été montré que les différents indices observent des fonctionnements différents vis-à-vis de l'obsolescence, et aussi de manière plus macro, dans une perspective discursive pour évaluer les combinaisons d'indices valables.

Nous avons mis en œuvre une méthodologie dont l'une des caractéristiques est d'être reproductible. Nous pensons que nous y sommes parvenue même s'il reste un certain nombre de points à améliorer, notamment l'intégration complète des outils OCAS et STAAT (dans la mesure du possible car l'utilisation de SPAD par exemple contraint le caractère reproductible de notre méthode) au sein de la plate-forme LINGUASTREAM.

Enfin, le prototype d'aide à la mise à jour de textes encyclopédiques mis en place a été évalué par des experts. Les résultats sont encourageants et laissent penser qu'on peut encore aller plus loin et améliorer le système de repérage des segments obsolescents si *(i)* la tâche est re-précisée, la notion de segment obsolescent redéfinie et affinée et *(ii)* le repérage des indices linguistiques et discursifs est repris et amélioré.

Index

Index

Bibliographie

Bibliographie

J. F. ALLEN : Towards a general theory of action and time. *Artificial Intelligence*, 23:p. 123–154, 1984.

D. BANKS : *Introduction à la linguistique systémique de l'anglais*. Paris, 2005. www.univ-brest.fr/erla/membres/banksdocs/Journee

S. BENHAZEZ, J.-P. DESCLÉS et J.-L. MINEL : Modèle d'exploration contextuelle pour l'analyse sémantique de textes. *In Actes de TALN 2001*, 2001.

E. BENVENISTE : *Problèmes de linguistique générale I*. Editions Gallimard, 1966.

J. BERRI, E. CARTIER, J.-P. DESCLÈS, A. JACKIEWICZ et J.-L. MINEL : Filtrage automatique de textes. Cargèse, 1996. TALN.

D. BIBER : *Variation across speech and writing*. Cambridge University Press, Cambridge, 1988.

D. BIBER : A typology of english texts. *Linguistics*, 27:3–43, 1989.

D. BIBER, U. CONNOR et T. ALBIN-UPTON : *Discourse on the move : using corpus analysis to describe discourse structure*. John Benjamins, 2007.

F. BILHAUT : Analyse automatique de la structure thématique du discours pour la naviguation documentaire. *In Workshop ATALA "Modéliser et décrire l'organisation discursive à l'heure du document numérique", Semaine du Document Numérique 2004*, La Rochelle, France, 2004.

F. BILHAUT : *Analyse automatique de structures thématiques discursives - Application à la recherche d'information*. Thèse de doctorat, Université de Caen, 2006.

F. BILHAUT : Analyse thématique automatique fondée sur la notion d'univers de discours. *Discours*, 1:en ligne, 2007.

A. BORILLO : Les adverbes de référence temporelle. *Communiversation*, 29, 1983. DRLAV, Paris.

A. BORILLO : Le lexique de l'espace : les noms et adjectifs de localisation interne. *Cahiers de grammaire*, 13:3–22, 1988.

A. BORILLO : A propos de la localisation spatiale. *Langue Française*, 86, 1990.

A. BORILLO : Les adjectifs et l'aspect en français. *Cahiers Chronos*, 2:177–189, 1998a.

A. BORILLO : *L'espace et son expression en français*. Ophrys, Paris, 1998b.

A. BORILLO : Les adverbes de temps dans la structuration du discours. mai 2003a. Colloque "L'adverbe", Arras.

A. BORILLO : Place et portée des adverbes de temps dans la structures de phrase et dans la structure de discours. mai 2003b. Colloque "L'adverbe", Arras.

A. BOUFFIER : *Analyse discursive automatique de textes - Application à la modélisation de textes incitatifs*. Thèse de doctorat, Université Paris Nord - Villetaneuse, 2008.

A. BOUHAFS : *Utilisation de la méthode d'exploration contextuelle pour une extraction d'informations sur le web dédiées à la veille*. Thèse de doctorat, Université paris IV - Sorbonne, 2005.

S. CARTER-THOMAS : *La cohérence textuelle*. XX, 2001.

E. CARTIER : Analyse automatique des textes : l'exemple des informations définitoires. Sfax, Tunisie, 1998. RIFRA'98.

P. CHARAUDEAU : Catégories de langue, catégories de discours et contrat de communication. *In Parcours Linguistique de discours spécialisés*, Editions Scientifiques Européennes, pages 315–325. Peter Lang S.A., Paris, 1992.

M. CHAROLLES : L'encadrement du discours, univers, champs, domaine et espaces. *Cahiers de Recherche linguistique*, 6, 1997.

M. CHAROLLES : Les adverbiaux cadratifs : fonction et classification. 2002a. URL `http://www.ltm.ens.fr/siteACFT/SiteLATTICEACFT4.htm`.

M. CHAROLLES : Le fonctionnement textuel des adverbiaux cadratifs. les adverbiaux temporels : un exemple. 2002b. URL `http://www.ltm.ens.fr/siteACFT/SiteLATTICEACFT4.htm`.

M. CHAROLLES : Les adverbiaux cadratifs et leur contribution à la cohésion du discours. avril 2003. Conférence à Toulouse.

M. CHAROLLES, A. LEDRAOULEC, M.-P. PÉRY-WOODLEY et L. SARDA : Temporal and spatial dimensions of discourse organization. *Journal of French Language Studies*, pages 115–130, 2005.

M. CHAROLLES et M.-P. PÉRY-WOODLEY : Les adverbiaux cadratifs - introduction. *Langue Française*, 148:pp. 3–8, 2005.

M. CHAROLLES et D. VIGIER : Les adverbiaux en position préverbale : portée cadrative et organisation des discours. *Langue Française*, 148:pp. 9–30, 2005.

A. CONDAMINES, J. REBEYROLLE et A. SOUBEILLE : Variation de la terminologie dans le temps : une méthode linguistique pour mesurer l'évolution de la connaissance en corpus. *In Actes Euralex International congress*, pages 547–557, Lorient, France, 2004.

J. COUTO, L. LUNDQUIST et J.-L. MINEL : Navigation interactive pour l'apprentissage en linguistique textuelle. *In Environnements Informatiques pour l'Apprentissage Humain*, Montpellier, 2005.

J. COUTO et J.-L. MINEL : Interfaces dynamiques de fouilles textuelles. *In RIAO 2004*, pages 420–430, Avignon, France, 2004.

A. CULIOLI : *Pour une linguistique de l'énonciation*, volume 2 - Formalisation et opérations de repérage. Ophrys, 1999.

B. DAILLE et E. MORIN : Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations. *T.A.L.*, 41:601–621, 2000.

L. DEGAND et T. SANDERS : The impact of relational markers on expository text comprehension in l1 and l2. *Reading and Writing*, 7-8(15):739–758, 2002.

J.-P. DESCLÈS : Systémes d'exploration contextuelle. *In Actes du colloque sur le calcul du sens et contexte*. Université de Caen, 1996.

J.-P. DESCLÈS et Z. GUENTCHEVA : Comment déterminer les significations du passé composé par une exploration contextuelle ? *Langue Française*, 138(1):48–60, 2003.

J.-P. DESCLÈS, Z. GUENTCHEVA, D. MAIRE-REPPERT et H.-G. OH : A propos de la catégorie grammaticale du temps et de l'aspect. *In Parcours Linguistique de discours spécialisés*, Editions Scientifiques Européennes, pages 291–299. Peter Lang S.A., Paris, 1992.

J.-P. DESCLÉS, C. JOUIS, H.-G. OH et D. M. REPPERT : Exploration contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte. *In* D. HERIN-AIME, R. DIENG, J-P. REGOURD et J.P. ANGOUJARD, éditeurs : *Knowledge modeling and expertise transfer*, pages pp.371–400, Amsterdam., 1991.

O. DUCROT et T. TODOROV : *Dictionnaire encyclopédique des sciences du langage*. Seuil, 1972.

M. EHRMANN : *Les entités nommées de la linguiistique au TAL : statut théorique et méthodes de désambiguisation*. Thèse de doctorat, Université Paris 7, 2008.

P. ENJALBERT : *Sémantique et TALN*. Hermes, 2005.

T. FAWCETT : Roc graphs : Notes and practical considerations for researchers. Rapport technique, HP Laboratories, 2003.

M. FAYOL : *Le récit et sa construction*. Delachaux et Niestlé, Lausanne, 1994.

O. FERRET, B. GRAU, J.-L. MINEL et S. PORHIEL : Repérage des structures thématiques dans des textes. TALN, juillet 2001.

D. GARCIA : *Analyse automatique des textes pour l'organisation causale des actions*. Thèse de doctorat, Université Paris-Sorbonne, 1998.

B. GOUJON : *Utilisation de l'exploration contextuelle pour l'aide à la veille technologique : Réalisation du système informatique VIGITEXT*. Thèse de doctorat, Université de Paris IV - Sorbonne, 2000. URL `http://www.lalic.paris4.sorbonne.fr/Theses/Goujon/`.

P. GUELPA : *Introduction à l'analyse linguistique*, chapitre La pragmatique. Quelques grands débats actuels, pages 208–214. Armaud Colin, 1997.

B. HABERT : *Instruments et ressources électroniques pour le français*. Ophrys, 2005.

M.A.K. HALLIDAY : *An introduction to functional grammar*. Edward Arnold, London, 1985.

M.A.K. HALLIDAY et R. HASAN : *Cohesion in English*. Longman Group Limited, London, 1976.

M. HEARST : Multi-paragraph segmentation of expository texts. *In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*, 1994.

N. HERNANDEZ : *Description et Détection Automatique de Structures de Texte*. Thèse de doctorat, Université de Paris XI, Décembre 2004.

N. HERNANDEZ et B. GRAU : Automatic extraction of meta-descriptors for text description. *In International Conference on Recent Advances In Natural Language Processing (RANLP)*, Borovets, Bulgaria, 2003.

M. HO-DAC et M. LAIGNELET : Temporal structure and thematic progression : a case study on french corpora. *In* M. AURNAGUE, M. BRAS, A. LEDRAOULEC et L. VIEU, éditeurs : *First International Symposium on the Exploration and Modelling of Meaning*, Biarritz, 2005.

M. HODAC : *La position initiale dans l'organisation du discours : une exploration en corpus*. Thèse de doctorat, Université de Toulouse 2 - Le Mirail, 2007.

M. HODAC et M.-P. PÉRY-WODLEY : Méthodologie exploratoire outillée pour l'étude de l'organisation du discours. *In Congrès Mondial de Linguistique Française*, Paris, France, 2008.

G. HRIPCSAK et D.F. HEITJAN : Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35(2):99–110, 2002.

F. IBEKWE-SANJUAN : Annotation d'indices de nouveautés dans les écrits scientifiques et techniques. *In Colloque Indice, Index, Indexation*, 2005.

P. IMBS : *L'emploi des temps verbaux en francais moderne*. Klinsieck, 1968.

A. JACKIEWICZ : *L'expression de la causalité dans les textes*. Thèse de doctorat, Université Paris-Sorbonne, 1998.

A. JACKIEWICZ : Repérage et délimitation des cadres organisationnels pour la segmentation automatique des textes. *In Actes de CIFT'02*, pages p. 95–107, Hammamet, Tunisie, 2002.

A. JACKIEWICZ : Les séries linéaires dans le discours. *Langue Française*, 148:pp. 95–110, 2005.

A. JACKIEWICZ et J.-L. MINEL : L'identification des structures discursives engendrées par les cadres organisationnels. Batz-sur-mer, juin 2003. TALN 2003.

M.-P. JACQUES et J. REBEYROLLE : Titres et structuration des documents. *In Actes du Colloque International Discours et Document*, pages 1–12, Caen, France, 2006.

R. JAKOBSON : Linguistique et poétique. *Essais de linguistique générale*, 1, 1963.

K. JOHNSON et H. JOHNSON : *Encyclopaedic dictionary of applied linguistics*, chapitre XX, pages 55–57, 99–101. Blackwell Publishers Ltd, Oxford, 1999.

K. Sparck JONES, S. WALKER et S.E. ROBERTSON : A probabilistic model of information retrieval : development and comparative experiments. *Information Processing and Management*, 36:Part 1 : p. 779–808 ; Part 2 : p. 809–840, 2000.

H. KAMP et C. ROHRER : *Meaning, use and interpretation of Language*, chapitre Tense in Texts, pages p. 250–269. De Gruyter, 1983.

M. LAIGNELET : Les titres et les cadres de discours temporels - structuration des discours et organisation de l'information. Mémoire de D.E.A., Université Toulouse-Le Mirail, 2004.

M. LAIGNELET : Analyse discursive pour le repérage automatique de segments d'information évolutive. *In Congrès de l'ACFAS - Colloque Description linguistique pour le traitement automatique du français, Montréal, Canada*, 2006a.

M. LAIGNELET : Repérage de segments d'information évolutive dans des documents de type encyclopédique. *In* Mertens P., Fairon C., Dister A. et Watrin P., éditeurs : *RECITAL'06 - Verbum ex machina - Actes de la 13e conférence sur le traitement automatique des langues, Jeunes Chercheurs*, volume 2, pages 690–699. Presses Universtaires de Louvain, 2006b.

M. LAIGNELET : Les titres et les introducteurs de cadre comme indices pour le repérage de segments d'information évolutive. *In International Symposium on Discourse and Document (ISDD)*, Caen, France, 2006c.

M. LAIGNELET : La recherche d'information évolutive dans des documents de type encyclopédique : l'apport de techniques linguistiques. *In Actes de la 4e conférence en recherche d'information et application - Rencontre de jeunes chercheurs en recherche d'information*, pages 449–454, 2007.

M. LAIGNELET et C. PIMM : Utiliser une segmentation thématique texttiling pour le repérage de segments d'information évolutive dans un corpus de textes encyclopédiques. *In RECITAL'07 - Actes de la 14e conférence sur le traitement automatique des langues, Jeunes Chercheurs*, pages 449–454. IRIT Presses, 2007.

M. LAIGNELET et F. RIOULT : Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. *In Actes de TALN 2009*, 2009. prix du "Best Paper".

A. LEDRAOULEC et M.-P. PÉRY-WOODLEY : Encadrement temporel et relations de discours. *Langue Française*, 148:pp. 45–60, 2005.

F. LEPRIOL : A data processing sequence to extract terms and semantic relation between terms. Brest, France, 1999. HCP'99.

F. LEPRIOL : *Extraction et capitalisation automatiques de connaissances à partir de documents textuels. SEEK-JAVA : identification et interprétation de relations entre concepts.* Thèse de doctorat, Université paris Sorbonne, 2000.

W. LI, J. HAN et J. PEI : Cmar : Accurate and efficient classification based on multiple class-association rules. *In IEEE International Conference on Data Mining*, 2001.

S. LOUDCHER-RABASEDA : *Contribution à l'extraction automatique de connaissances : application à l'analyse clinique de la marche*. Thèse de doctorat, Université Lyon 1, 1996.

C. LUC : *Représentation et composition des structures visuelles et rhétoriques du texte*. Thèse de doctorat, Université Paul Sabatier, 2000.

C. LUC et J. VIRBEL : Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum*, XXIII(1):104–123, 2001.

N. LUCAS et B. CRÉMILLEUX : Fouille de textes hiérarchisés appliquée à la détection de fautes. *Document Numérique*, 3(8):107–133, 2003.

D. MALRIEU et F. RASTIER : Genres et variations morphosyntaxiques. *Traitement Automatiqe des langues*, 42(2):548–577, 2001.

I. MANI : *Automatic summarization*. John Benjamins Publishing Compagny, Amsterdam/Philadelphie, 2001.

D. MARCU : *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, 2000.

R. MARTIN : Les univers de croyance dans la théorie sémantique. *Langage et croyance*, Bruxelles, Mardaga, 1987.

L. MASCHERIN : *Analyse morphosémantique de l'aspectuo-temporalité en français. Le cas du préfixe re-*. Thèse de doctorat, Université de Nancy 2, 2008.

J.-L. MINEL : *Filtrage sémantique, du résumé automatique à la fouille de textes*. Hermès, 2002.

G. MOURAD : Présentation de connaissances linguistiques pour le repérage et l'extraction de citations. *In Actes de TALN (RECITAL'2000)*, pages pp. 495–501, Lausanne, 2000.

G. MOURAD : La segmentation de textes par exploration contextuelle automatique, présentation du module segatex. *In Actes du colloque "Inscription Spatiale du Langage : structure et processus" (ISLsp)*, 2002.

P.-A. MULLER et N. GAERTNER : *Modélisation objet avec UML*. Eyrolles, 2005.

A. NAZARENKO : *Sémantique et Corpus*, chapitre Sur quelles sémantiques reposent les méthodes automatiques d'accés au contenu textuel ?, pages 211–244. Hermès, 2005.

E. PASCUAL et M.-P. PÉRY-WOODLEY : La définition dans le texte. *In* J.-L. NESPOULOUS et J. VIRBEL, éditeurs : *Textes de type consigne - Perception, action, cognition*, pages 65–88. Toulouse, 1995.

M.-P. PÉRY-WOODLEY : Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle. Rapport technique, Carnet de grammaire, Rapport n8, 2000. thèse d'habilitation.

A. PICTON : Combining clues to explore knowledge evolution. *In Actes de la Conférence internationale Terminology and Knowledge Engineering (TKE)*, Copenhague, Danemark, 2008.

S. PORHIEL : Les séquences thématiques. *Langue Française*, 148:pp. 111–126, 2005.

F. RASTIER : *La linguistique de corpus*, chapitre Enjeux épistémologiques de la linguistique de corpus. Presses Universitaires de Rennes, 2005.

J. REBEYROLLE : Forme linguistique et fonction discursive des titres de sections. 2004. URL `www.univ-tlse2.fr/erss/membres/hodac/VIZU/`. document de travail.

REICHENBACH : *Elements of symbolic logic New- york*. Free-Press,, 1966. 1ère édition 1947.

T. REINHART : Conditions for text coherence. *Poetics Today*, 1(4):pp. 161–180, 1980.

M. RIEGEL, J.-C. PELLAT et R. RIOUL : *Grammaire méthodique du français*. PUF, 1994.

F. RIOULT, B. ZANUTTINI et B. CRÉMILLEUX : Apport de la négation pour la classification supervisée à l'aide d'associations. *In Conférence d'Apprentissage*, pages 183–196, 2008.

G. SALTON, A. WONG et C.S. YANG : A vector space model for automatic indexing. *In Commun. ACM*, volume 18, pages p. 613–620, 1975.

L. SARDA : Fonctionnement des cadres spatiaux dans les résumés de film. *Langue Française*, 148:pp. 61–99, 2005.

G.-E. SARFATI : *Eléments d'analyse du discours*. Nathan, Paris, 1997.

L. SINI : *Les marqueurs linguistiques de la présence de l'auteur*, chapitre Le cas de « tout de même pas », marqueur d'empathie, et de « justement », marqueur méta-énonciatif, pages 19–33. L'Harmattan, 2005.

A. TADROS : *Prediction in text*, volume 10. English language research, University of Birmingham, Birmingham, 1985.

S. TEUFEL : Meta-discourse markers and problem-structuring in scientific articles. *In Workshop on Discourse Structure and Discourse Markers*, Montreal, 1998. ACL 1998.

S. TEUFEL : *Argumentative Zoning*. Thèse de doctorat, Université de Edimbourg, 1999.

S. TEUFEL, J. CARLETTA et M. MOENS : An annotation scheme for discourse-level argumentation in research articles. *In Proceedings of EACL*, 1999.

S. TEUFEL et M. MOENS : Summarizing scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4), 2002.

E. TOGNINI-BONELLI : *Corpus Linguistics at work*. John Benjamins Publishing Company, 2001.

Y. TOUSSAINT : Extraction de connaissances à partir de textes structurés. *Document numérique*, 8:11–34, 2004.

S. TUFFÉRY : *Data Mining et statistique décisionnelle - l'intelligence des données*. Editions Technip, 2007.

M. VALETTE et N. GRABAR : Caractérisation de textes à contenu idéologique : statistique textuelle ou extracation de syntagme ? l'exemple du projet princip. *In Actes des 7èmes Journées internationalesD'Analyse statistique des Données textuelles (JADT 2004)*, 2004.

Co VET : *Temps, aspects et adverbes de temps en français contemporain*. Thèse de doctorat, Université d'Amsterdam, 1980.

C. VETTERS et E. SKIBINSKA : Le futur : une question de temps ou de mode ? remarques générales et analyse du "présent-futur" perfectif polonais. *Cahiers Chronos*, 2:247–266, 1998.

D. VIGIER : Les syntagmes prépositionnels en « en n » détachés en tête de phrase référant à des domaines d'activité. *In Grammaires et Lexiques Comparés*, volume 21, pages 97–122, Bari, Monopoli, Italie, 2003.

R. VINOT, N. GRABAR et M. VALETTE : Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'internet. *In Actes des TALN 2003*, 2003.

J. VIRBEL : Langage et métalangage dans le texte du point de vue de l'édition en informatique textuelle. *Cahiers de grammaire*, 10:5–72, 1985.

J. VIRBEL : Eléments d'analyse du titre. *Inscription Spatiale du Langage : structures et processus*, pages 123–132, 2002.

M. VUILLAUME : *L'énonciation dans tous ses états*, chapitre La temporalité discursive, pages 433–449. Peter lang, 2008.

H. WEINRICH : *Le temps*. Editions du Seuil, Paris, 1973.

A. WIDLÖCHER : *Analyse macro-sémantique des structures rhétoriques du discours. Cadre théorique et modèle opératoire.* Thèse de doctorat, Université de Caen, 2008.

A. WIDLÖCHER et F. BILHAUT : La plate-forme linguastream : un outil d'exploration linguistique sur corpus. *In Actes de la 12e Conférence Traitement Automatique du Langage Naturel (TALN)*, Dourdan, France, 2005.

N. ZERIDA, N. LUCAS et B. CRÉMILLEUX : Combinaison de descripteurs linguistiques et de structure pour la fouille d'articles biomédicaux. *In Schedae*, pages 69–78, 2006.

Annexes

Annexe A

Explication de la notation UML

Il est impossible de donner une représentation graphique complète d'un logiciel, ou de tout autre système complexe, de même qu'il est impossible de représenter entièrement une statue (à trois dimensions) par des photographies (à deux dimensions). Mais il est possible de donner sur un tel système des vues partielles, analogues chacune à une photographie d'une statue, et dont la conjonction donnera une idée utilisable en pratique sans risque d'erreur grave.

UML n'est pas une méthode (*i.e.* une description normative des étapes de modélisation). C'est un langage graphique qui permet de représenter et de communiquer les divers aspects d'un système d'information. Aux graphiques sont associés des textes qui expliquent leur contenu (Muller et Gaertner, 2005).

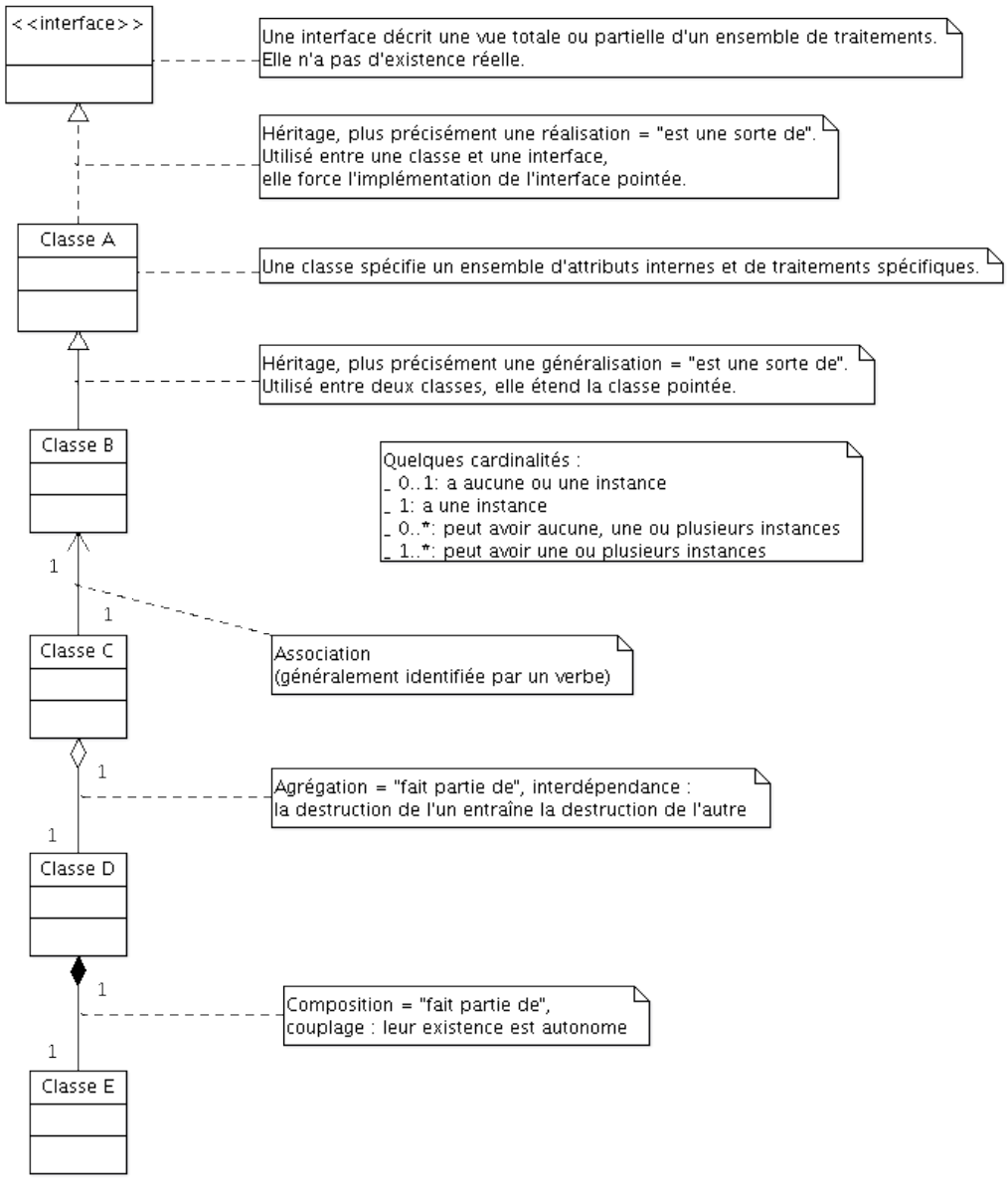


Schéma A.1 - *Rappel des notations UML*

Annexe B

Transformation des données textuelles en données structurées

B.1 Description de la base de données créée

Le schéma ci-dessous rend compte du schéma relationnel de données de la base `discourse.sql` créée en sortie de l'outil OCAS.

ANALYSISUNIT (id, type, order, precedingId*, text) ENCLOSINGUNIT (id, type, order, enclosedAnalysisUnitId*) ENCLOSEDUNIT (id, type, order, enclosingAnalysisUnitId*) FEATURE (name, value, analysisUnitId*, enclosingUnitId*, enclosedUnitId*) DYNAMICUNIT (*analysisUnitId*, *name*, quantity)

La table `analysisUnit` récupère les informations concernant le segment textuel sur lequel se baseront les analyses. Pour notre étude, il s'agit des unités `sentence` (les phrases) et des unités `title` (les titres de nos corpus).

La table `enclosedUnit` contient les indices discursifs qui sont contenus à l'intérieur des `analysisUnit`. Il peut s'agir par exemple de l'indice `advTemp` ou de l'indice `argum`.

La table `enclosingUnit` contient les marqueurs discursifs qui englobent les `analysisUnit`. Il peut s'agir par exemple de l'indice `parag` ou de l'indice `zone`.

Pour chacune de ces trois tables, nous importons quatre types de données issues des analyses faites à l'aide de LINGUASTREAM :

– `id` : correspond à l'ordre d'apparition réel et linéaire du marqueur (indice ou appear) ou de l'élément, tous types confondus. Chaque numéro est unique.

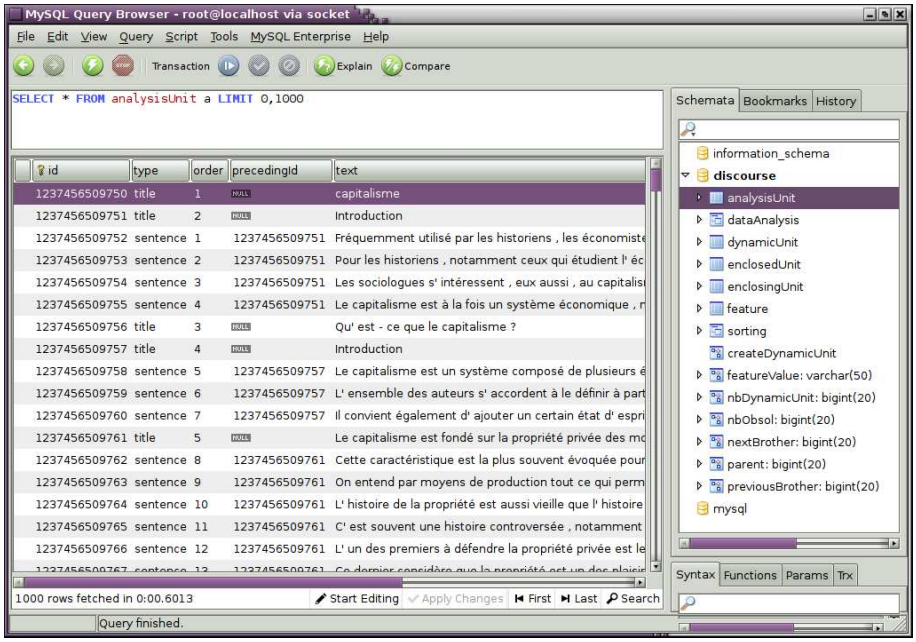


Figure B.1 - *Image de la table analysisUnit*

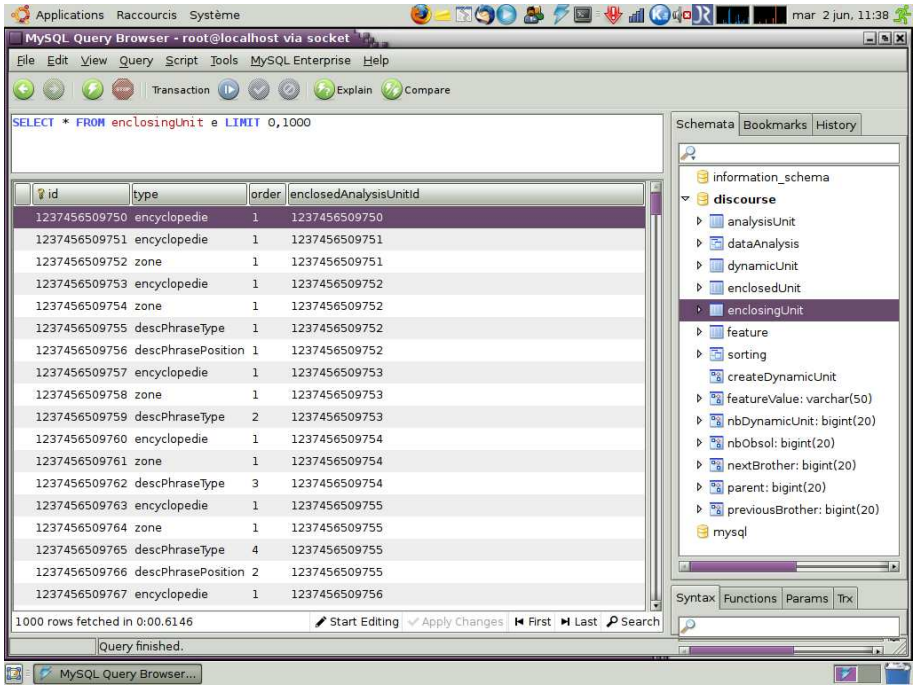


Figure B.2 - *Image de la table enclosingUnit*

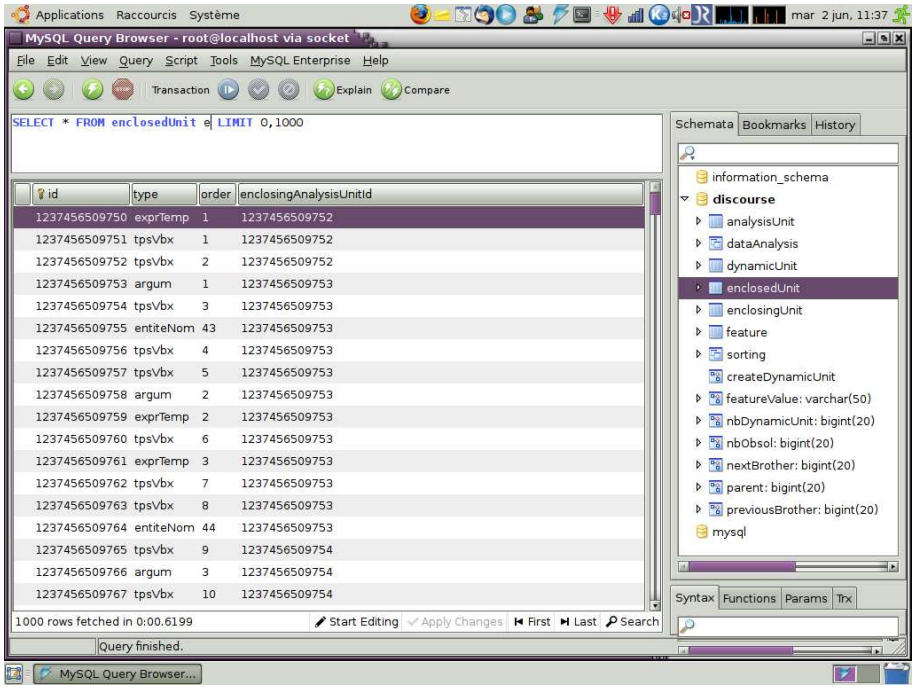


Figure B.3 - *Image de la table enclosedUnit*

Les `id` des trois tables sont du type clés primaires ;

- `type` : donne le nom du marqueur (enclosedUnit ou enclosingUnit) ou de l'élément ;
- `order` : correspond au numéro du marqueur (enclosedUnit ou enclosingUnit) ou de l'élément donné par les modules Linguastream. Contrairement à `id`, ces numéros dépendent du type et ne sont donc pas uniques ;
- `elementId` : cet identifiant permet, pour les tables `enclosingUnit` et `enclosedUnit\` de relier, grâce à des clés étrangères (`indiceToElement` et `appearToElement`) la relation entre le marqueur discursif et l'élément dans lequel il apparaît. En ce qui concerne la table `analysisUnit`, `elementId` permet, pour chaque élément `sentence`, de le relier avec un élément `title`, et ce à l'aide de la clé étrangère `elementToElement`. Pour les éléments `title`, cette information reste vide (valeur : $NULL$. Les `elementId` des trois tables sont du type clés étrangères.

En ce qui concerne la table `feature`, elle contient toutes les structures de traits pour chacun des éléments et marqueurs des tables que nous venons de présenter (`analysisUnit`, `enclosingUnit`, et `enclosedUnit`). Elle est l'étape intermédiaire vers la construction dynamique des noms de variables tels qu'ils ap-

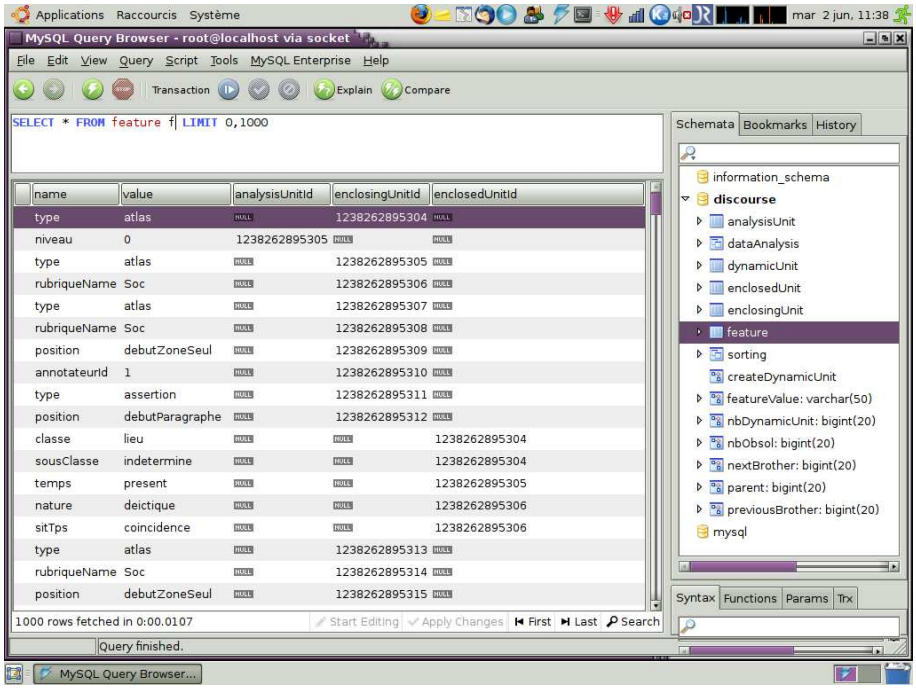


Figure B.4 - *Image de la table feature*

paraîtront dans la vue finale.

Les types de données composant cette table `feature` sont :

- `name` : correspond à la première partie d'un trait pour une structure de traits donnée ;
- `value` : correspond à la deuxième partie d'un trait pour une structure de traits donnée ;
- `elementId` : fait référence à l'élément sur lequel la structure de trait s'applique à l'aide de la clé étrangère `featureToElement` ;
- `appearId` : idem mais s'il s'agit d'un marqueur de type `enclosingUnit` à l'aide de la clé étrangère `featureToElement` ;
- `indiceId` : idem mais s'il s'agit d'un marqueur de type `enclosedUnit` à l'aide de la clé étrangère `featureToElement`.

S'il y a plusieurs traits dans la structure de traits, alors seront créées autant de lignes que de traits. Le fait qu'il y ait plusieurs valeurs identiques pour les attributs `elementId`, `appearId` et `indiceId` nous renseignent sur l'unité de la structure de traits en question.

Enfin, la table `dynamicUnit` finalise le nommage dynamique des variables à partir des unités discursives effectivement repérées et annotées dans le corpus.

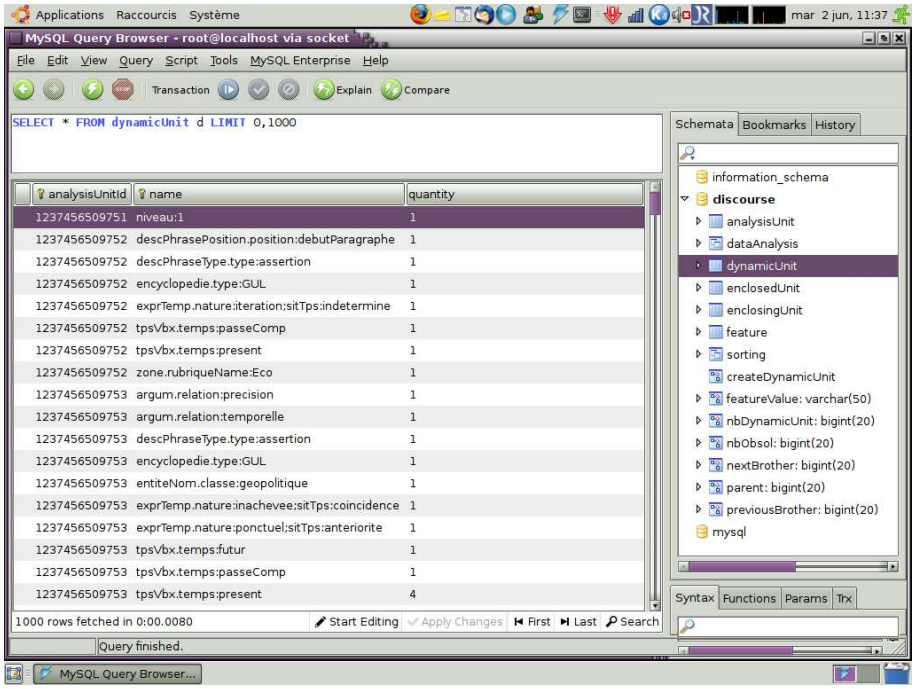


Figure B.5 - *Image de la table dynamicUnit*

Le but est à la fois de concaténer les noms des indices (tables `analysisUnit`, `enclosingUnit` et `enclosedUnit`) avec les traits de la table `feature` correspondante et de faire le décompte des occurrences par unité d'analyse.

- `analysisUnitId` : correspond à l'identifiant de l'unité d'analyse ;
- `name` : le nom de la variable est composé du nom de l'unité discursive auquel on concatène sa structure de traits. Par exemple, un adverbial temporel de `nature` ; *ponctuel* et de `sitTps` *coincidence* sera transformé en $advTemp.nature : ponctuel; sitTps : coincidence$;
- `quantity` : indique la quantité d'occurrences pour la variable donnée.

La version MySql que nous utilisons ne permet pas de transformer des éléments en ligne (ce qui nous intéresse ici, les noms de variables) en noms de colonne. Nous sommes alors passé par un programme JAVA qui permet cette transformation et la création des vues dont nous avons parlé dans la section 6.3 (vues $dataAnalysis$ et $Sorting$).

Ce système de tables nous permet de conserver un maximum d'informations et surtout de *mettre à plat* des informations textuelles et discursives très diverses et de niveaux différents. Il nous est maintenant possible, à partir de ces tables de créer n'importe quelle vue. Comme nous l'avons déjà expliqué, une vue est une table

virtuelle qui nous permet de rassembler des informations provenant de plusieurs tables différentes comme ce sera le cas pour cette étude.

B.2 Exemple de transformation des données

B.2.1 Texte original

Agriculture et environnement

§ L'application du progrès technique, là où elle se réalise, conduit à l'existence d'une agriculture intensive, qui utilise beaucoup de produits chimiques et de machines, tout en n'ayant plus besoin que d'un petit nombre d'agriculteurs. Cet emploi quelquefois massif de moyens venant de l'extérieur des exploitations peut avoir des effets néfastes sur l'environnement. [...]

Source : Corpus GLI (fiche Géographie - Agriculture)

Exemple B.6 - *Un exemple de transformation des données*

B.2.2 Texte annoté manuellement et automatiquement par l'outil ALIDIS en sortie de LINGUASTREAM

```
<titre niveau="1">
  <ls:b type="title" id="65" layer="92">
  <ls:b type="entiteNom" id="24" layer="25"> Agriculture </ls:b>
  et environnement </ls:b>
</titre>

<paragraphe>
  <ls:a type="premierParag" id="18" layer="90" anchor="start"/>
  <ls:a type="descPhraseType" id="344" layer="77" anchor="start"/>
  <ls:a type="descPhrasePosition" id="157" layer="77" anchor="start"/>

   <ls:a type="sentence" id="347" layer="75" anchor="start"/>
   L'application du <ls:a type="ptVue" id="50" layer="46" anchor="start"/>
   progrès <ls:a type="ptVue" id="50" layer="46" anchor="end"/>
   technique,là où elle se <ls:b type="tpsVbx" id="599" layer="25">
   réalise </ls:b>, <ls:b type="tpsVbx" id="600" layer="25">
   conduit </ls:b> à l'existence d'une
   <ls:a type="entiteNom" id="515" layer="25" anchor="start"/>
   agriculture <ls:a type="entiteNom" id="515" layer="25" anchor="end"/>
   intensive, qui  <ls:b type="tpsVbx" id="601" layer="25">
   utilise </ls:b> beaucoup de produits chimiques, de machines, tout en
   n'ayant plus besoin que d'un petit
   <ls:a type="position" id="131" layer="50" anchor="start"/>
   <ls:a type="entiteNom" id="516" layer="25" anchor="start"/>
   nombre  <ls:a type="preposition" id="1341" layer="13" anchor="start"/>
   d'agriculteurs <ls:a type="entiteNom" id="516" layer="25" anchor="end"/>
   <ls:a type="position" id="131" layer="50" anchor="end"/> .
   <ls:a type="sentence" id="347" layer="75" anchor="end"/>
   <ls:a type="descPhrasePosition" id="157" layer="77" anchor="end"/>
   <ls:a type="descPhraseType" id="344" layer="77" anchor="end"/>

   <ls:a type="descPhraseType" id="345" layer="77" anchor="start"/>
   <ls:a type="descPhrasePosition" id="158" layer="77" anchor="start"/>
   <ls:a type="sentence" id="348" layer="75" anchor="start"/>
   Cet emploi quelquefois massif de moyens venant de l'extérieur des
   exploitations <ls:b type="tpsVbx" id="602" layer="25"> peut </ls:b>
   avoir des effets néfastes sur l'environnement.
   <ls:a type="sentence" id="348" layer="75" anchor="end"/>
   <ls:a type="descPhrasePosition" id="158" layer="77" anchor="end"/>
```

```
<ls:a type="descPhraseType" id="345" layer="77" anchor="end"/>
<ls:a type="premierParag" id="18" layer="90" anchor="end"/>
</paragraphe>
```

nb. : nous avons volontairement simplifié la vue de la sortie XML en ne conservant que les indices discursifs pertienents pour notre travail.

B.2.3 Insertion dans la base de données

PREMIÈRE PHRASE :

ANALYSISUNIT (1239527722351, sentence, 347, 1239527722350*, "L'application du progrès technique, là où elle se réalise, conduit à l'existence d'une agriculture intensive, qui utilise beaucoup de produits chimiques et de machines, tout en n'ayant plus besoin que d'un petit nombre d'agriculteurs.")

ENCLOSINGUNIT (1239527723393, encyclopedie, 1, 1239527722351*)
ENCLOSINGUNIT (1239527723394, zone, 14, 1239527722351*)
ENCLOSINGUNIT (1239527723395, premierParag, 18, 1239527722351*)
ENCLOSINGUNIT (1239527723396, descPhraseType, 344, 1239527722351*)
ENCLOSINGUNIT (1239527723397, descPhrasePosition, 157, 1239527722351*)

ENCLOSEDUNIT (1239527723552, ptVue, 50, 1239527722351*)
ENCLOSEDUNIT (1239527723553, tpsVbx, 599, 1239527722351*)
ENCLOSEDUNIT (1239527723554, tpsVbx, 600, 1239527722351*)
ENCLOSEDUNIT (1239527723555, entitenom, 515, 1239527722351*)
ENCLOSEDUNIT (1239527723556, tpsVbx, 601, 1239527722351*)
ENCLOSEDUNIT (1239527723557, position, 131, 1239527722351*)
ENCLOSEDUNIT (1239527723558, entiteNom, 516, 1239527722351*)

FEATURE (type, null, analysisUnitId*, 1239527723393*, enclosedUnitId*)
FEATURE (rubriqueName, null, analysisUnitId*, 1239527723394*, enclosedUnitId*)
FEATURE (position, debutDivisionSeul, analysisUnitId*, 1239527723395*, enclosedUnitId*)
FEATURE (type, assertion, analysisUnitId*, 1239527723396*, enclosedUnitId*)
FEATURE (position, debutParagraphe, analysisUnitId*, 1239527723397*, enclosedUnitId*)

FEATURE (type, prevision, analysisUnitId*, enclosingUnitId*, 1239527723552*)
FEATURE (temps, present, analysisUnitId*, enclosingUnitId*, 1239527723553*)
FEATURE (temps, present, analysisUnitId*, enclosingUnitId*, 1239527723554*)
FEATURE (classe geopolitique, 1, analysisUnitId*, enclosingUnitId*, 1239527723555*)
FEATURE (temps, present, analysisUnitId*, enclosingUnitId*, 1239527723556*)
FEATURE (typeIndice, entiteNom, analysisUnitId*, enclosingUnitId*, 1239527723557*)
FEATURE (typePos, END, analysisUnitId*, enclosingUnitId*, 1239527723557*)

FEATURE (classe, geopolitique, analysisUnitId*, enclosingUnitId*, 1239527723558*)

SECONDE PHRASE :

ANALYSISUNIT (1239527722352, sentence, 348, 1239527722350*, "Cet emploi quelquefois massif de moyens venant de l'extérieur des exploitations peut avoir des effets néfastes sur l'environnement.")

ENCLOSINGUNIT (1239527723393, encyclopedie, 1, 1239527722351*)
ENCLOSINGUNIT (1239527723394, zone, 14, 1239527722351*)
ENCLOSINGUNIT (1239527723395, premierParag, 18, 1239527722351*)
ENCLOSINGUNIT (1239527723396, descPhraseType, 345, 1239527722351*)
ENCLOSINGUNIT (1239527723397, descPhrasePosition, 158, 1239527722351*)

ENCLOSEDUNIT (1239527723559, tpsVbx, 602, 1239527722352*)

FEATURE (type, null, analysisUnitId*, 1239527723393*, enclosedUnitId*)
FEATURE (rubriqueName, null, analysisUnitId*, 1239527723394*, enclosedUnitId*)
FEATURE (position, debutDivisionSeul, analysisUnitId*, 1239527723395*, enclosedUnitId*)
FEATURE (type, assertion, analysisUnitId*, 1239527723396*, enclosedUnitId*)
FEATURE (position, debutParagraphe, analysisUnitId*, 1239527723397*, enclosedUnitId*)

FEATURE (temps, present, analysisUnitId*, enclosingUnitId*, 1239527723559*)

TITRE :

ANALYSISUNIT (123952772235O, title, 65, NULL*, "Agriculture et environnement")

ENCLOSINGUNIT (1239527723391, encyclopedie, 1, 1239527722350*)
ENCLOSINGUNIT (1239527723392, zone, 14, 1239527722350*)

ENCLOSEDUNIT (1239527723551, entiteNom, 24, 1239527722350*)

FEATURE (niveau, 1, 1239527722350*, enclosingUnitId*, enclosedUnitId*)

FEATURE (type, null, analysisUnitId*, 1239527723391*, enclosedUnitId*)
FEATURE (rubriqueName, null, analysisUnitId*, 1239527723392*, enclosedUnitId*)

FEATURE (classe, geopolitique, analysisUnitId*, enclosingUnitId*, 1239527723551*)

DYNAMICUNIT :

```
<row>
      <field name="analysisUnitId">1239527722351</field>
      <field name="name">descPhrasePosition.position:debutParagraphe</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722351</field>
      <field name="name">descPhraseType.type:assertion</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722351</field>
      <field name="name">encyclopedie.type:</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722351</field>
      <field name="name">entiteNom.classe:geopolitique</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722351</field>
      <field name="name">position.typeIndice:entiteNom;typePos:END</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722351</field>
      <field name="name">premierParag.position:debutDivisionSeul</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722351</field>
      <field name="name">ptVue.type:prevision</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722351</field>
      <field name="name">tpsVbx.temps:present</field>
      <field name="quantity">3</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722351</field>
      <field name="name">zone.rubriqueName:</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722352</field>
      <field name="name">descPhrasePosition.position:finParagraphe</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722352</field>
      <field name="name">descPhraseType.type:assertion</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722352</field>
      <field name="name">encyclopedie.type:</field>
      <field name="quantity">1</field>
    </row>
```

```
    <row>
      <field name="analysisUnitId">1239527722352</field>
      <field name="name">premierParag.position:debutDivisionSeul</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722352</field>
      <field name="name">tpsVbx.temps:present</field>
      <field name="quantity">1</field>
    </row>
    <row>
      <field name="analysisUnitId">1239527722352</field>
      <field name="name">zone.rubriqueName:</field>
      <field name="quantity">1</field>
    </row>
```

B.2.4 Sortie *sorting* : pour l'apprentissage automatique

```
<row>
      <field name="sentenceId">1239527722351</field>
      <field name="obsol">0</field>
      <field name="argum.relation:correction">0</field>
      <field name="argum.relation:identite">0</field>
      <field name="descPhrasePosition.position:seuleInParagraphe">0</field>
      <field name="descPhraseType.type:assertion">1</field>
      <field name="encyclopedie.type:">1</field>
      <field name="premierParag.position:debutZoneSeul">0</field>
      <field name="tpsVbx.temps:present">3</field>
      <field name="zone.rubriqueName:Med">0</field>
      <field name="descPhrasePosition.position:debutParagraphe">1</field>
      <field name="entiteNom.classe:mesure;sousClasse:indetermine">0</field>
      <field name="premierParag.position:debutDivision">0</field>
      <field name="entiteNom.classe:plus">0</field>
      <field name="descPhrasePosition.position:finParagraphe">0</field>
      <field name="entiteNom.classe:geopolitique">1</field>
      <field name="exprTemp.nature:ponctuel;sitTps:trespasse">0</field>
      <field name="entiteNom.classe:sigle">0</field>
      <field name="position.typeIndice:entiteNom;typePos:IC">0</field>
      <field name="ptVue.type:restriction">0</field>
      <field name="tpsVbx.temps:passeComp">0</field>
      <field name="argum.relation:consecution">0</field>
      <field name="argum.relation:liste">0</field>
      <field name="position.typeIndice:ptVue;typePos:IC">0</field>
      <field name="ptVue.type:distance">0</field>
      <field name="dernierParag.position:finZone">0</field>
      <field name="entiteNom.classe:lieu;sousClasse:riviere">0</field>
      <field name="premierParag.position:debutZone">0</field>
      <field name="tpsVbx.temps:passeAnt">0</field>
      <field name="zone.rubriqueName:Soc">0</field>
      <field name="exprTemp.nature:deictique;sitTps:coincidence">0</field>
      <field name="exprTemp.nature:inachevee;sitTps:trespasse">0</field>
      <field name="exprTemp.nature:ponctuel;sitTps:indetermine">0</field>
      <field name="tpsVbx.temps:passeSimple">0</field>
      <field name="exprTemp.nature:inachevee;sitTps:anteriorite">0</field>
      <field name="exprTemp.nature:ponctuel;sitTps:anteriorite">0</field>
      <field name="ptVue.type:recence">0</field>
      <field name="position.typeIndice:exprTemp;typePos:END">0</field>
      <field name="entiteNom.classe:mesure;sousClasse:geopolitique">0</field>
```

```
  <field name="entiteNom.classe:personne">0</field>
  <field name="exprTemp.nature:duree;sitTps:indetermine">0</field>
  <field name="position.typeIndice:entiteNom;typePos:AMORCE">0</field>
  <field name="zone.rubriqueName:ScTechn">0</field>
  <field name="entiteNom.classe:mesure;sousClasse:evolutif">0</field>
  <field name="ptVue.type:definition">0</field>
  <field name="exprTemp.nature:iteration;sitTps:indetermine">0</field>
  <field name="argum.relation:explication">0</field>
  <field name="entiteNom.classe:lieu;sousClasse:indetermine">0</field>
  <field name="position.typeIndice:entiteNom;typePos:END">1</field>
  <field name="argum.relation:temporelle">0</field>
  <field name="tpsVbx.temps:imparfait">0</field>
  <field name="argum.relation:illustration">0</field>
  <field name="argum.relation:opposition">0</field>
  <field name="tpsVbx.temps:conditionnel">0</field>
  <field name="dernierParag.position:finDivision">0</field>
  <field name="entiteNom.classe:mesure;sousClasse:fixe">0</field>
  <field name="entiteNom.classe:lieu;sousClasse:pointCardinal">0</field>
  <field name="exprTemp.nature:ponctuel;sitTps:coincidence">0</field>
  <field name="position.typeIndice:ptVue;typePos:END">0</field>
  <field name="zone.rubriqueName:ArtLitt">0</field>
  <field name="tpsVbx.temps:plusQuePft">0</field>
  <field name="entiteNom.classe:lieu;sousClasse:ville">0</field>
  <field name="exprTemp.nature:anaphorique;sitTps:indetermine">0</field>
  <field name="tpsVbx.temps:futur">0</field>
  <field name="entiteNom.classe:lieu;sousClasse:pays">0</field>
  <field name="position.typeIndice:argum;typePos:IC">0</field>
  <field name="ptVue.type:jugement">0</field>
  <field name="argum.relation:justification">0</field>
  <field name="exprTemp.nature:deictique;sitTps:posteriorite">0</field>
  <field name="ptVue.type:prevision">1</field>
  <field name="argum.relation:precision">0</field>
  <field name="ptVue.type:source">0</field>
  <field name="position.typeIndice:ptVue;typePos:AMORCE">0</field>
  <field name="zone.rubriqueName:Geo">0</field>
  <field name="zone.rubriqueName:Hist">0</field>
  <field name="position.typeIndice:exprTemp;typePos:IC">0</field>
  <field name="tpsVbx.temps:futAnt">0</field>
  <field name="zone.rubriqueName:">1</field>
  <field name="exprTemp.nature:inachevee;sitTps:coincidence">0</field>
  <field name="periVbs.accomplissement:debut">0</field>
  <field name="exprTemp.nature:deictique;sitTps:trespasse">0</field>
  <field name="exprTemp.nature:anaphorique;sitTps:coincidence">0</field>
  <field name="exprTemp.nature:deictique;sitTps:anteriorite">0</field>
  <field name="premierParag.position:debutDivisionSeul">1</field>
  <field name="ptVue.type:importance">0</field>
  <field name="periVbs.accomplissement:fin">0</field>
  <field name="title-&gt;niveau:0">0</field>
  <field name="title-&gt;niveau:1">1</field>
  <field name="title-&gt;entiteNom.classe:geopolitique">1</field>
  <field name="title-&gt;tpsVbx.temps:present">0</field>
  <field name="title-&gt;niveau:2">0</field>
  <field name="title-&gt;entiteNom.classe:sigle">0</field>
  <field name="title-&gt;ptVue.type:prevision">0</field>
  <field name="title-&gt;entiteNom.classe:lieu;sousClasse:indetermine">0</field>
  <field name="title-&gt;entiteNom.classe:lieu;sousClasse:pays">0</field>
  <field name="title-&gt;exprTemp.nature:ponctuel;sitTps:coincidence">0</field>
  <field name="title-&gt;exprTemp.nature:ponctuel;sitTps:anteriorite">0</field>
  <field name="title-&gt;ptVue.type:recence">0</field>
  <field name="title-&gt;entiteNom.classe:personne">0</field>
</row>
<row>
```

```
<field name="sentenceId">1239527722352</field>
<field name="obsol">0</field>
<field name="argum.relation:correction">0</field>
<field name="argum.relation:identite">0</field>
<field name="descPhrasePosition.position:seuleInParagraphe">0</field>
<field name="descPhraseType.type:assertion">1</field>
<field name="encyclopedie.type:">1</field>
<field name="premierParag.position:debutZoneSeul">0</field>
<field name="tpsVbx.temps:present">1</field>
<field name="zone.rubriqueName:Med">0</field>
<field name="descPhrasePosition.position:debutParagraphe">0</field>
<field name="entiteNom.classe:mesure;sousClasse:indetermine">0</field>
<field name="premierParag.position:debutDivision">0</field>
<field name="entiteNom.classe:plus">0</field>
<field name="descPhrasePosition.position:finParagraphe">1</field>
<field name="entiteNom.classe:geopolitique">0</field>
<field name="exprTemp.nature:ponctuel;sitTps:trespasse">0</field>
<field name="entiteNom.classe:sigle">0</field>
<field name="position.typeIndice:entiteNom;typePos:IC">0</field>
<field name="ptVue.type:restriction">0</field>
<field name="tpsVbx.temps:passeComp">0</field>
<field name="argum.relation:consecution">0</field>
<field name="argum.relation:liste">0</field>
<field name="position.typeIndice:ptVue;typePos:IC">0</field>
<field name="ptVue.type:distance">0</field>
<field name="dernierParag.position:finZone">0</field>
<field name="entiteNom.classe:lieu;sousClasse:riviere">0</field>
<field name="premierParag.position:debutZone">0</field>
<field name="tpsVbx.temps:passeAnt">0</field>
<field name="zone.rubriqueName:Soc">0</field>
<field name="exprTemp.nature:deictique;sitTps:coincidence">0</field>
<field name="exprTemp.nature:inachevee;sitTps:trespasse">0</field>
<field name="exprTemp.nature:ponctuel;sitTps:indetermine">0</field>
<field name="tpsVbx.temps:passeSimple">0</field>
<field name="exprTemp.nature:inachevee;sitTps:anteriorite">0</field>
<field name="exprTemp.nature:ponctuel;sitTps:anteriorite">0</field>
<field name="ptVue.type:recence">0</field>
<field name="position.typeIndice:exprTemp;typePos:END">0</field>
<field name="entiteNom.classe:mesure;sousClasse:geopolitique">0</field>
<field name="entiteNom.classe:personne">0</field>
<field name="exprTemp.nature:duree;sitTps:indetermine">0</field>
<field name="position.typeIndice:entiteNom;typePos:AMORCE">0</field>
<field name="zone.rubriqueName:ScTechn">0</field>
<field name="entiteNom.classe:mesure;sousClasse:evolutif">0</field>
<field name="ptVue.type:definition">0</field>
<field name="exprTemp.nature:iteration;sitTps:indetermine">0</field>
<field name="argum.relation:explication">0</field>
<field name="entiteNom.classe:lieu;sousClasse:indetermine">0</field>
<field name="position.typeIndice:entiteNom;typePos:END">0</field>
<field name="argum.relation:temporelle">0</field>
<field name="tpsVbx.temps:imparfait">0</field>
<field name="argum.relation:illustration">0</field>
<field name="argum.relation:opposition">0</field>
<field name="tpsVbx.temps:conditionnel">0</field>
<field name="dernierParag.position:finDivision">0</field>
<field name="entiteNom.classe:mesure;sousClasse:fixe">0</field>
<field name="entiteNom.classe:lieu;sousClasse:pointCardinal">0</field>
<field name="exprTemp.nature:ponctuel;sitTps:coincidence">0</field>
<field name="position.typeIndice:ptVue;typePos:END">0</field>
<field name="zone.rubriqueName:ArtLitt">0</field>
<field name="tpsVbx.temps:plusQuePft">0</field>
<field name="entiteNom.classe:lieu;sousClasse:ville">0</field>
```

```
      <field name="exprTemp.nature:anaphorique;sitTps:indetermine">0</field>
      <field name="tpsVbx.temps:futur">0</field>
      <field name="entiteNom.classe:lieu;sousClasse:pays">0</field>
      <field name="position.typeIndice:argum;typePos:IC">0</field>
      <field name="ptVue.type:jugement">0</field>
      <field name="argum.relation:justification">0</field>
      <field name="exprTemp.nature:deictique;sitTps:posteriorite">0</field>
      <field name="ptVue.type:prevision">0</field>
      <field name="argum.relation:precision">0</field>
      <field name="ptVue.type:source">0</field>
      <field name="position.typeIndice:ptVue;typePos:AMORCE">0</field>
      <field name="zone.rubriqueName:Geo">0</field>
      <field name="zone.rubriqueName:Hist">0</field>
      <field name="position.typeIndice:exprTemp;typePos:IC">0</field>
      <field name="tpsVbx.temps:futAnt">0</field>
      <field name="zone.rubriqueName:">1</field>
      <field name="exprTemp.nature:inachevee;sitTps:coincidence">0</field>
      <field name="periVbs.accomplissement:debut">0</field>
      <field name="exprTemp.nature:deictique;sitTps:trespasse">0</field>
      <field name="exprTemp.nature:anaphorique;sitTps:coincidence">0</field>
      <field name="exprTemp.nature:deictique;sitTps:anteriorite">0</field>
      <field name="premierParag.position:debutDivisionSeul">1</field>
      <field name="ptVue.type:importance">0</field>
      <field name="periVbs.accomplissement:fin">0</field>
      <field name="title-&gt;niveau:0">0</field>
      <field name="title-&gt;niveau:1">1</field>
      <field name="title-&gt;entiteNom.classe:geopolitique">1</field>
      <field name="title-&gt;tpsVbx.temps:present">0</field>
      <field name="title-&gt;niveau:2">0</field>
      <field name="title-&gt;entiteNom.classe:sigle">0</field>
      <field name="title-&gt;ptVue.type:prevision">0</field>
      <field name="title-&gt;entiteNom.classe:lieu;sousClasse:indetermine">0</field>
      <field name="title-&gt;entiteNom.classe:lieu;sousClasse:pays">0</field>
      <field name="title-&gt;exprTemp.nature:ponctuel;sitTps:coincidence">0</field>
      <field name="title-&gt;exprTemp.nature:ponctuel;sitTps:anteriorite">0</field>
      <field name="title-&gt;ptVue.type:recence">0</field>
      <field name="title-&gt;entiteNom.classe:personne">0</field>
    </row>
```

B.2.5 Sortie *dataAnalysis* : pour les statistiques descriptives et l'ACP

```
<row>
      <field name="Id">1239527722351</field>
      <field name="Text">L' application du progrès technique, là où elle se réalise,
       conduit à l existence d'une agriculture intensive, qui utilise
        beaucoup de produits chimiques (engrais, pesticides) et de machines,
        tout en n'ayant plus besoin que d'un petit nombre d'agriculteurs .</field>
      <field name="Obs">0</field>
      <field name="V001-argum.relation:correction">0</field>
      <field name="V002-argum.relation:identite">0</field>
      <field name="V003-descPhrasePosition.position:seuleInParagraphe">0</field>
      <field name="V004-descPhraseType.type:assertion">1</field>
      <field name="V005-encyclopedie.type:">1</field>
      <field name="V006-premierParag.position:debutZoneSeul">0</field>
      <field name="V007-tpsVbx.temps:present">3</field>
      <field name="V008-zone.rubriqueName:Med">0</field>
      <field name="V009-descPhrasePosition.position:debutParagraphe">1</field>
      <field name="V010-entiteNom.classe:mesure;sousClasse:indetermine">0</field>
      <field name="V011-premierParag.position:debutDivision">0</field>
```

```
<field name="V012-entiteNom.classe:plus">0</field>
<field name="V013-descPhrasePosition.position:finParagraphe">0</field>
<field name="V014-entiteNom.classe:geopolitique">1</field>
<field name="V015-exprTemp.nature:ponctuel;sitTps:trespasse">0</field>
<field name="V016-entiteNom.classe:sigle">0</field>
<field name="V017-position.typeIndice:entiteNom;typePos:IC">0</field>
<field name="V018-ptVue.type:restriction">0</field>
<field name="V019-tpsVbx.temps:passeComp">0</field>
<field name="V020-argum.relation:consecution">0</field>
<field name="V021-argum.relation:liste">0</field>
<field name="V022-position.typeIndice:ptVue;typePos:IC">0</field>
<field name="V023-ptVue.type:distance">0</field>
<field name="V024-dernierParag.position:finZone">0</field>
<field name="V025-entiteNom.classe:lieu;sousClasse:riviere">0</field>
<field name="V026-premierParag.position:debutZone">0</field>
<field name="V027-tpsVbx.temps:passeAnt">0</field>
<field name="V028-zone.rubriqueName:Soc">0</field>
<field name="V029-exprTemp.nature:deictique;sitTps:coincidence">0</field>
<field name="V030-exprTemp.nature:inachevee;sitTps:trespasse">0</field>
<field name="V031-exprTemp.nature:ponctuel;sitTps:indetermine">0</field>
<field name="V032-tpsVbx.temps:passeSimple">0</field>
<field name="V033-exprTemp.nature:inachevee;sitTps:anteriorite">0</field>
<field name="V034-exprTemp.nature:ponctuel;sitTps:anteriorite">0</field>
<field name="V035-ptVue.type:recence">0</field>
<field name="V036-position.typeIndice:exprTemp;typePos:END">0</field>
<field name="V037-entiteNom.classe:mesure;sousClasse:geopolitique">0</field>
<field name="V038-entiteNom.classe:personne">0</field>
<field name="V039-exprTemp.nature:duree;sitTps:indetermine">0</field>
<field name="V040-position.typeIndice:entiteNom;typePos:AMORCE">0</field>
<field name="V041-zone.rubriqueName:ScTechn">0</field>
<field name="V042-entiteNom.classe:mesure;sousClasse:evolutif">0</field>
<field name="V043-ptVue.type:definition">0</field>
<field name="V044-exprTemp.nature:iteration;sitTps:indetermine">0</field>
<field name="V045-argum.relation:explication">0</field>
<field name="V046-entiteNom.classe:lieu;sousClasse:indetermine">0</field>
<field name="V047-position.typeIndice:entiteNom;typePos:END">1</field>
<field name="V048-argum.relation:temporelle">0</field>
<field name="V049-tpsVbx.temps:imparfait">0</field>
<field name="V050-argum.relation:illustration">0</field>
<field name="V051-argum.relation:opposition">0</field>
<field name="V052-tpsVbx.temps:conditionnel">0</field>
<field name="V053-dernierParag.position:finDivision">0</field>
<field name="V054-entiteNom.classe:mesure;sousClasse:fixe">0</field>
<field name="V055-entiteNom.classe:lieu;sousClasse:pointCardinal">0</field>
<field name="V056-exprTemp.nature:ponctuel;sitTps:coincidence">0</field>
<field name="V057-position.typeIndice:ptVue;typePos:END">0</field>
<field name="V058-zone.rubriqueName:ArtLitt">0</field>
<field name="V059-tpsVbx.temps:plusQuePft">0</field>
<field name="V060-entiteNom.classe:lieu;sousClasse:ville">0</field>
<field name="V061-exprTemp.nature:anaphorique;sitTps:indetermine">0</field>
<field name="V062-tpsVbx.temps:futur">0</field>
<field name="V063-entiteNom.classe:lieu;sousClasse:pays">0</field>
<field name="V064-position.typeIndice:argum;typePos:IC">0</field>
<field name="V065-ptVue.type:jugement">0</field>
<field name="V066-argum.relation:justification">0</field>
<field name="V067-exprTemp.nature:deictique;sitTps:posteriorite">0</field>
<field name="V068-ptVue.type:prevision">1</field>
<field name="V069-argum.relation:precision">0</field>
<field name="V070-ptVue.type:source">0</field>
<field name="V071-position.typeIndice:ptVue;typePos:AMORCE">0</field>
<field name="V072-zone.rubriqueName:Geo">0</field>
<field name="V073-zone.rubriqueName:Hist">0</field>
```

```
    <field name="V074-position.typeIndice:exprTemp;typePos:IC">0</field>
    <field name="V075-tpsVbx.temps:futAnt">0</field>
    <field name="V076-zone.rubriqueName:">1</field>
    <field name="V077-exprTemp.nature:inachevee;sitTps:coincidence">0</field>
    <field name="V078-periVbs.accomplissement:debut">0</field>
    <field name="V079-exprTemp.nature:deictique;sitTps:trespasse">0</field>
    <field name="V080-exprTemp.nature:anaphorique;sitTps:coincidence">0</field>
    <field name="V081-exprTemp.nature:deictique;sitTps:anteriorite">0</field>
    <field name="V082-premierParag.position:debutDivisionSeul">1</field>
    <field name="V083-ptVue.type:importance">0</field>
    <field name="V084-periVbs.accomplissement:fin">0</field>
    <field name="V085-title-&gt;niveau:0">0</field>
    <field name="V086-title-&gt;niveau:1">1</field>
    <field name="V087-title-&gt;entiteNom.classe:geopolitique">1</field>
    <field name="V088-title-&gt;tpsVbx.temps:present">0</field>
    <field name="V089-title-&gt;niveau:2">0</field>
    <field name="V090-title-&gt;entiteNom.classe:sigle">0</field>
    <field name="V091-title-&gt;ptVue.type:prevision">0</field>
    <field name="V092-title-&gt;entiteNom.classe:lieu;sousClasse:indetermine">0</field>
    <field name="V093-title-&gt;entiteNom.classe:lieu;sousClasse:pays">0</field>
    <field name="V094-title-&gt;exprTemp.nature:ponctuel;sitTps:coincidence">0</field>
    <field name="V095-title-&gt;exprTemp.nature:ponctuel;sitTps:anteriorite">0</field>
    <field name="V096-title-&gt;ptVue.type:recence">0</field>
    <field name="V097-title-&gt;entiteNom.classe:personne">0</field>
  </row>
  <row>
    <field name="Id">1239527722352</field>
    <field name="Text">Cet emploi quelquefois massif de moyens venant de
    l'extérieur des exploitations peut avoir des effets néfastes sur
    l'environnement.</field>
    <field name="Obs">0</field>
    <field name="V001-argum.relation:correction">0</field>
    <field name="V002-argum.relation:identite">0</field>
    <field name="V003-descPhrasePosition.position:seuleInParagraphe">0</field>
    <field name="V004-descPhraseType.type:assertion">1</field>
    <field name="V005-encyclopedie.type:">1</field>
    <field name="V006-premierParag.position:debutZoneSeul">0</field>
    <field name="V007-tpsVbx.temps:present">1</field>
    <field name="V008-zone.rubriqueName:Med">0</field>
    <field name="V009-descPhrasePosition.position:debutParagraphe">0</field>
    <field name="V010-entiteNom.classe:mesure;sousClasse:indetermine">0</field>
    <field name="V011-premierParag.position:debutDivision">0</field>
    <field name="V012-entiteNom.classe:plus">0</field>
    <field name="V013-descPhrasePosition.position:finParagraphe">1</field>
    <field name="V014-entiteNom.classe:geopolitique">0</field>
    <field name="V015-exprTemp.nature:ponctuel;sitTps:trespasse">0</field>
    <field name="V016-entiteNom.classe:sigle">0</field>
    <field name="V017-position.typeIndice:entiteNom;typePos:IC">0</field>
    <field name="V018-ptVue.type:restriction">0</field>
    <field name="V019-tpsVbx.temps:passeComp">0</field>
    <field name="V020-argum.relation:consecution">0</field>
    <field name="V021-argum.relation:liste">0</field>
    <field name="V022-position.typeIndice:ptVue;typePos:IC">0</field>
    <field name="V023-ptVue.type:distance">0</field>
    <field name="V024-dernierParag.position:finZone">0</field>
    <field name="V025-entiteNom.classe:lieu;sousClasse:riviere">0</field>
    <field name="V026-premierParag.position:debutZone">0</field>
    <field name="V027-tpsVbx.temps:passeAnt">0</field>
    <field name="V028-zone.rubriqueName:Soc">0</field>
    <field name="V029-exprTemp.nature:deictique;sitTps:coincidence">0</field>
    <field name="V030-exprTemp.nature:inachevee;sitTps:trespasse">0</field>
    <field name="V031-exprTemp.nature:ponctuel;sitTps:indetermine">0</field>
```

```
<field name="V032-tpsVbx.temps:passeSimple">0</field>
<field name="V033-exprTemp.nature:inachevee;sitTps:anteriorite">0</field>
<field name="V034-exprTemp.nature:ponctuel;sitTps:anteriorite">0</field>
<field name="V035-ptVue.type:recence">0</field>
<field name="V036-position.typeIndice:exprTemp;typePos:END">0</field>
<field name="V037-entiteNom.classe:mesure;sousClasse:geopolitique">0</field>
<field name="V038-entiteNom.classe:personne">0</field>
<field name="V039-exprTemp.nature:duree;sitTps:indetermine">0</field>
<field name="V040-position.typeIndice:entiteNom;typePos:AMORCE">0</field>
<field name="V041-zone.rubriqueName:ScTechn">0</field>
<field name="V042-entiteNom.classe:mesure;sousClasse:evolutif">0</field>
<field name="V043-ptVue.type:definition">0</field>
<field name="V044-exprTemp.nature:iteration;sitTps:indetermine">0</field>
<field name="V045-argum.relation:explication">0</field>
<field name="V046-entiteNom.classe:lieu;sousClasse:indetermine">0</field>
<field name="V047-position.typeIndice:entiteNom;typePos:END">0</field>
<field name="V048-argum.relation:temporelle">0</field>
<field name="V049-tpsVbx.temps:imparfait">0</field>
<field name="V050-argum.relation:illustration">0</field>
<field name="V051-argum.relation:opposition">0</field>
<field name="V052-tpsVbx.temps:conditionnel">0</field>
<field name="V053-dernierParag.position:finDivision">0</field>
<field name="V054-entiteNom.classe:mesure;sousClasse:fixe">0</field>
<field name="V055-entiteNom.classe:lieu;sousClasse:pointCardinal">0</field>
<field name="V056-exprTemp.nature:ponctuel;sitTps:coincidence">0</field>
<field name="V057-position.typeIndice:ptVue;typePos:END">0</field>
<field name="V058-zone.rubriqueName:ArtLitt">0</field>
<field name="V059-tpsVbx.temps:plusQuePft">0</field>
<field name="V060-entiteNom.classe:lieu;sousClasse:ville">0</field>
<field name="V061-exprTemp.nature:anaphorique;sitTps:indetermine">0</field>
<field name="V062-tpsVbx.temps:futur">0</field>
<field name="V063-entiteNom.classe:lieu;sousClasse:pays">0</field>
<field name="V064-position.typeIndice:argum;typePos:IC">0</field>
<field name="V065-ptVue.type:jugement">0</field>
<field name="V066-argum.relation:justification">0</field>
<field name="V067-exprTemp.nature:deictique;sitTps:posteriorite">0</field>
<field name="V068-ptVue.type:prevision">0</field>
<field name="V069-argum.relation:precision">0</field>
<field name="V070-ptVue.type:source">0</field>
<field name="V071-position.typeIndice:ptVue;typePos:AMORCE">0</field>
<field name="V072-zone.rubriqueName:Geo">0</field>
<field name="V073-zone.rubriqueName:Hist">0</field>
<field name="V074-position.typeIndice:exprTemp;typePos:IC">0</field>
<field name="V075-tpsVbx.temps:futAnt">0</field>
<field name="V076-zone.rubriqueName:">1</field>
<field name="V077-exprTemp.nature:inachevee;sitTps:coincidence">0</field>
<field name="V078-periVbs.accomplissement:debut">0</field>
<field name="V079-exprTemp.nature:deictique;sitTps:trespasse">0</field>
<field name="V080-exprTemp.nature:anaphorique;sitTps:coincidence">0</field>
<field name="V081-exprTemp.nature:deictique;sitTps:anteriorite">0</field>
<field name="V082-premierParag.position:debutDivisionSeul">1</field>
<field name="V083-ptVue.type:importance">0</field>
<field name="V084-periVbs.accomplissement:fin">0</field>
<field name="V085-title-&gt;niveau:0">0</field>
<field name="V086-title-&gt;niveau:1">1</field>
<field name="V087-title-&gt;entiteNom.classe:geopolitique">1</field>
<field name="V088-title-&gt;tpsVbx.temps:present">0</field>
<field name="V089-title-&gt;niveau:2">0</field>
<field name="V090-title-&gt;entiteNom.classe:sigle">0</field>
<field name="V091-title-&gt;ptVue.type:prevision">0</field>
<field name="V092-title-&gt;entiteNom.classe:lieu;sousClasse:indetermine">0</field>
<field name="V093-title-&gt;entiteNom.classe:lieu;sousClasse:pays">0</field>
```

```
    <field name="V094-title-&gt;exprTemp.nature:ponctuel;sitTps:coincidence">0</field>
    <field name="V095-title-&gt;exprTemp.nature:ponctuel;sitTps:anteriorite">0</field>
    <field name="V096-title-&gt;ptVue.type:recence">0</field>
    <field name="V097-title-&gt;entiteNom.classe:personne">0</field>
  </row>
```

Annexe C

Résultats SPAD

C.1 Libellées des variables : correspondance codes et noms des variables

```
SELECTION DES INDIVIDUS ET DES VARIABLES UTILES
VARIABLES CONTINUES ACTIVES
   143 VARIABLES
-------------------------------------------------------------------------------------------------------------
   1. Obs                                                          ( CONTINUE )
   2. V001-descPhrasePosition.position:debutParagraphe             ( CONTINUE )
   3. V002-descPhraseType.type:assertion                           ( CONTINUE )
   5. V004-exprTemp.nature:iteration;sitTps:indetermine            ( CONTINUE )
   6. V005-tpsVbx.temps:passeComp                                  ( CONTINUE )
   7. V006-tpsVbx.temps:present                                    ( CONTINUE )
   8. V007-zone.rubriqueName:Eco                                   ( CONTINUE )
   9. V008-argum.relation:precision                                ( CONTINUE )
  10. V009-argum.relation:temporelle                               ( CONTINUE )
  11. V010-entiteNom.classe:geopolitique                           ( CONTINUE )
  12. V011-exprTemp.nature:inachevee;sitTps:coincidence            ( CONTINUE )
  13. V012-exprTemp.nature:ponctuel;sitTps:anteriorite             ( CONTINUE )
  14. V013-tpsVbx.temps:futur                                      ( CONTINUE )
  15. V014-argum.relation:identite                                 ( CONTINUE )
  16. V015-argum.relation:correction                               ( CONTINUE )
  17. V016-descPhrasePosition.position:finParagraphe               ( CONTINUE )
  18. V017-argum.relation:consecution                              ( CONTINUE )
  19. V018-argum.relation:explication                              ( CONTINUE )
  20. V019-entiteNom.classe:personne                               ( CONTINUE )
  21. V020-entiteNom.classe:sigle                                  ( CONTINUE )
  22. V021-entiteNom.classe:lieu;sousClasse:indetermine            ( CONTINUE )
  23. V022-position.typeIndice:entiteNom;typePos:IC                ( CONTINUE )
  24. V023-exprTemp.nature:ponctuel;sitTps:trespasse               ( CONTINUE )
  25. V024-tpsVbx.temps:plusQuePft                                 ( CONTINUE )
  26. V025-tpsVbx.temps:imparfait                                  ( CONTINUE )
  27. V026-entiteNom.classe:plus                                   ( CONTINUE )
  28. V027-tpsVbx.temps:conditionnel                               ( CONTINUE )
  29. V028-exprTemp.nature:ponctuel;sitTps:coincidence             ( CONTINUE )
  30. V029-ptVue.type:prevision                                    ( CONTINUE )
  31. V030-exprTemp.nature:deictique;sitTps:trespasse              ( CONTINUE )
  32. V031-ptVue.type:distance                                     ( CONTINUE )
  33. V032-position.typeIndice:argum;typePos:IC                    ( CONTINUE )
  34. V033-argum.relation:illustration                             ( CONTINUE )
  35. V034-exprTemp.nature:deictique;sitTps:coincidence            ( CONTINUE )
  36. V035-descPhrasePosition.position:seuleInParagraphe           ( CONTINUE )
  37. V036-descPhraseType.type:interrogation                       ( CONTINUE )
  38. V037-periVbs.accomplissement:debut                           ( CONTINUE )
  39. V038-ptVue.type:recence                                      ( CONTINUE )
  40. V039-argum.relation:justification                            ( CONTINUE )
  41. V040-tpsVbx.temps:passeSimple                                ( CONTINUE )
  42. V041-position.typeIndice:ptVue;typePos:END                   ( CONTINUE )
  43. V042-dernierParag.position:finDivision                        ( CONTINUE )
  44. V043-exprTemp.nature:inachevee;sitTps:anteriorite            ( CONTINUE )
  45. V044-position.typeIndice:exprTemp;typePos:END                ( CONTINUE )
  46. V045-entiteNom.classe:lieu;sousClasse:pays                   ( CONTINUE )
  47. V046-argum.relation:liste                                    ( CONTINUE )
  48. V047-position.typeIndice:entiteNom;typePos:END               ( CONTINUE )
  49. V048-position.typeIndice:ptVue;typePos:IC                    ( CONTINUE )
```

```
 50. V049-exprTemp.nature:inachevee;sitTps:trespasse              ( CONTINUE )
 51. V050-ptVue.type:restriction                                  ( CONTINUE )
 52. V051-argum.relation:concession                               ( CONTINUE )
 53. V052-position.typeIndice:exprTemp;typePos:IC                 ( CONTINUE )
 54. V053-premierParag.position:debutDivision                     ( CONTINUE )
 55. V054-exprTemp.nature:ponctuel;sitTps:indetermine             ( CONTINUE )
 56. V055-dernierParag.position:finZone                           ( CONTINUE )
 57. V056-entiteNom.classe:mesure;sousClasse:indetermine          ( CONTINUE )
 58. V057-tpsVbx.temps:futAnt                                     ( CONTINUE )
 59. V058-argum.relation:opposition                               ( CONTINUE )
 60. V062-entiteNom.classe:lieu;sousClasse:ville                  ( CONTINUE )
 61. V063-exprTemp.nature:anaphorique;sitTps:indetermine          ( CONTINUE )
 62. V064-exprTemp.nature:duree;sitTps:indetermine                ( CONTINUE )
 63. V065-ptVue.type:jugement                                     ( CONTINUE )
 64. V066-premierParag.position:debutDivisionSeul                 ( CONTINUE )
 65. V067-zone.rubriqueName:Droit                                 ( CONTINUE )
 66. V068-zone.rubriqueName:Hist                                  ( CONTINUE )
 67. V069-exprTemp.nature:anaphorique;sitTps:coincidence          ( CONTINUE )
 68. V070-exprTemp.nature:anaphorique;sitTps:posteriorite         ( CONTINUE )
 69. V071-exprTemp.nature:duree;sitTps:coincidence                ( CONTINUE )
 70. V072-tpsVbx.temps:passeAnt                                   ( CONTINUE )
 71. V073-argum.relation:incertitude                              ( CONTINUE )
 72. V074-argum.relation:resume                                   ( CONTINUE )
 73. V076-position.typeIndice:entiteNom;typePos:AMORCE            ( CONTINUE )
 74. V077-entiteNom.classe:mesure;sousClasse:evolutif             ( CONTINUE )
 75. V078-entiteNom.classe:mesure;sousClasse:geopolitique         ( CONTINUE )
 76. V079-tpsVbx.temps:condPasse                                  ( CONTINUE )
 77. V080-position.typeIndice:ptVue;typePos:AMORCE                ( CONTINUE )
 78. V081-ptVue.type:source                                       ( CONTINUE )
 79. V082-exprTemp.nature:deictique;sitTps:anteriorite            ( CONTINUE )
 80. V083-entiteNom.classe:lieu;sousClasse:riviere                ( CONTINUE )
 81. V084-ptVue.type:definition                                   ( CONTINUE )
 82. V085-zone.rubriqueName:Soc                                   ( CONTINUE )
 83. V086-entiteNom.classe:mesure;sousClasse:fixe                 ( CONTINUE )
 84. V087-periVbs.accomplissement:fin                             ( CONTINUE )
 85. V088-exprTemp.nature:ponctuel;sitTps:posteriorite            ( CONTINUE )
 86. V089-entiteNom.classe:lieu;sousClasse:adresse                ( CONTINUE )
 87. V090-periVbs.accomplissement:deroulement                     ( CONTINUE )
 88. V091-premierParag.position:debutZone                         ( CONTINUE )
 89. V092-zone.rubriqueName:FauneFlore                            ( CONTINUE )
 90. V093-exprTemp.nature:deictique;sitTps:posteriorite           ( CONTINUE )
 91. V094-zone.rubriqueName:Geo                                   ( CONTINUE )
 92. V095-entiteNom.classe:lieu;sousClasse:pointCardinal          ( CONTINUE )
 93. V096-entiteNom.classe:mesure;sousClasse:estimation           ( CONTINUE )
 94. V097-exprTemp.nature:deictique;sitTps:indetermine            ( CONTINUE )
 95. V098-ptVue.type:importance                                   ( CONTINUE )
 96. V099-zone.rubriqueName:ScTechn                               ( CONTINUE )
 97. V100-premierParag.position:debutZoneSeul                     ( CONTINUE )
 98. V101-zone.rubriqueName:Med                                   ( CONTINUE )
 99. V102-exprTemp.nature:inachevee;sitTps:indetermine            ( CONTINUE )
101. V104-position.typeIndice:argum;typePos:END                   ( CONTINUE )
102. V105-exprTemp.nature:duree;sitTps:trespasse                  ( CONTINUE )
103. V106-position.typeIndice:exprTemp;typePos:AMORCE             ( CONTINUE )
104. V107-exprTemp.nature:anaphorique;sitTps:anteriorite          ( CONTINUE )
106. V109-entiteNom.classe:moins                                  ( CONTINUE )
107. V110-zone.rubriqueName:                                      ( CONTINUE )
108. V111-zone.rubriqueName:Sport                                 ( CONTINUE )
109. V112-descPhraseType.type:exclamation                         ( CONTINUE )
110. V113-exprTemp.nature:inachevee;sitTps:posteriorite           ( CONTINUE )
111. V114-ptVue.type:enonciatif                                   ( CONTINUE )
112. V115-exprTemp.nature:ponctuel                                ( CONTINUE )
113. V116-zone.rubriqueName:ArtLitt                               ( CONTINUE )
114. V117-title->niveau:1                                         ( CONTINUE )
115. V118-title->tpsVbx.temps:present                             ( CONTINUE )
116. V119-title->niveau:2                                         ( CONTINUE )
117. V120-title->entiteNom.classe:geopolitique                    ( CONTINUE )
118. V121-title->tpsVbx.temps:passeComp                           ( CONTINUE )
119. V122-title->argum.relation:temporelle                        ( CONTINUE )
120. V123-title->exprTemp.nature:deictique;sitTps:coincidence     ( CONTINUE )
121. V124-title->exprTemp.nature:ponctuel;sitTps:anteriorite      ( CONTINUE )
122. V125-title->exprTemp.nature:ponctuel;sitTps:coincidence      ( CONTINUE )
123. V126-title->ptVue.type:distance                              ( CONTINUE )
124. V127-title->tpsVbx.temps:conditionnel                        ( CONTINUE )
125. V128-title->entiteNom.classe:personne                        ( CONTINUE )
126. V129-title->niveau:3                                         ( CONTINUE )
127. V130-title->ptVue.type:recence                               ( CONTINUE )
128. V131-title->entiteNom.classe:lieu;sousClasse:indetermine     ( CONTINUE )
129. V132-title->niveau:4                                         ( CONTINUE )
130. V133-title->niveau:5                                         ( CONTINUE )
131. V134-title->ptVue.type:prevision                             ( CONTINUE )
132. V135-title->entiteNom.classe:sigle                           ( CONTINUE )
133. V136-title->niveau:0                                         ( CONTINUE )
134. V137-title->entiteNom.classe:lieu;sousClasse:pays            ( CONTINUE )
135. V138-title->exprTemp.nature:ponctuel;sitTps:trespasse        ( CONTINUE )
```

```
136. V139-title->entiteNom.classe:lieu;sousClasse:ville           ( CONTINUE )
137. V140-title->tpsVbx.temps:imparfait                           ( CONTINUE )
138. V141-title->entiteNom.classe:lieu;sousClasse:riviere         ( CONTINUE )
139. V142-title->position.typeIndice:entiteNom;typePos:AMORCE     ( CONTINUE )
140. V143-title->tpsVbx.temps:passeSimple                         ( CONTINUE )
141. V144-title->exprTemp.nature:inachevee;sitTps:trespasse       ( CONTINUE )
142. V145-title->tpsVbx.temps:futur                               ( CONTINUE )
143. V146-title->exprTemp.nature:deictique;sitTps:posteriorite    ( CONTINUE )
144. V147-title->argum.relation:correction                        ( CONTINUE )
145. V148-title->ptVue.type:importance                            ( CONTINUE )
146. V149-title->exprTemp.nature:inachevee;sitTps:anteriorite     ( CONTINUE )
```

C.2 Statistiques de base

C.2.1 Corrélation entre les variables continues et la variable : Obs

- `V.TEST` = valeur-test associée à la probabilité du test de corrélation nulle
- `PROBA.` = probabilité associée au test de corrélation
- `CORR.` = corrélation entre la variable à caractériser et chacune des variables continues. Il faut noter que la corrélation est ici calculée après élimination des données manquantes
- `NUM. LIBELLE DE LÀ VARIABLE` = libellé complet des variables continues caractérisantes
- `POIDS` = effectif sur lequel est calculée la corrélation si le poids des individus est uniforme et le poids des individus si les individus ont des poids différents

V.TEST	PROBA.	CORR.	NUM. LIBELLE DE LÀ VARIABLE	POIDS
681.57	0.000	1.000	1. Obs	9916.000
26.00	0.000	0.255	117. V120-title->entiteNom.classe:geopolitique	9916.000
24.39	0.000	0.240	74. V077-entiteNom.classe:mesure;sousClasse:evolutif	9916.000
23.86	0.000	0.235	11. V010-entiteNom.classe:geopolitique	9916.000
19.33	0.000	0.192	91. V094-zone.rubriqueName:Geo	9916.000
18.91	0.000	0.188	35. V034-exprTemp.nature:deictique;sitTps:coincidence	9916.000
15.27	0.000	0.152	75. V078-entiteNom.classe:mesure;sousClasse:geopolitique	9916.000
9.94	0.000	0.099	129. V132-title->niveau:4	9916.000
8.53	0.000	0.085	29. V028-exprTemp.nature:ponctuel;sitTps:coincidence	9916.000
7.89	0.000	0.079	48. V047-position.typeIndice:entiteNom;typePos:END	9916.000
7.83	0.000	0.078	72. V074-argum.relation:resume	9916.000
7.83	0.000	0.078	56. V055-dernierParag.position:finZone	9916.000
6.63	0.000	0.066	93. V096-entiteNom.classe:mesure;sousClasse:estimation	9916.000
6.38	0.000	0.064	65. V067-zone.rubriqueName:Droit	9916.000
6.30	0.000	0.063	23. V022-position.typeIndice:entiteNom;typePos:IC	9916.000
6.09	0.000	0.061	17. V016-descPhrasePosition.position:finParagraphe	9916.000
5.85	0.000	0.059	28. V027-tpsVbx.temps:conditionnel	9916.000
5.80	0.000	0.058	78. V081-ptVue.type:source	9916.000
5.63	0.000	0.056	87. V090-periVbs.accomplissement:deroulement	9916.000
5.52	0.000	0.055	22. V021-entiteNom.classe:lieu;sousClasse:indetermine	9916.000
5.22	0.000	0.052	43. V042-dernierParag.position:finDivision	9916.000
5.10	0.000	0.051	46. V045-entiteNom.classe:lieu;sousClasse:pays	9916.000
4.73	0.000	0.047	27. V026-entiteNom.classe:plus	9916.000
4.68	0.000	0.047	85. V088-exprTemp.nature:ponctuel;sitTps:posteriorite	9916.000
4.39	0.000	0.044	79. V082-exprTemp.nature:deictique;sitTps:anteriorite	9916.000
4.23	0.000	0.042	12. V011-exprTemp.nature:inachevee;sitTps:coincidence	9916.000
4.05	0.000	0.041	30. V029-ptVue.type:prevision	9916.000
3.73	0.000	0.037	143. V146-title->exprTemp.nature:deictique;sitTps:posteriorite	9916.000
3.70	0.000	0.037	133. V136-title->niveau:0	9916.000
3.62	0.000	0.036	21. V020-entiteNom.classe:sigle	9916.000
3.45	0.000	0.035	73. V076-position.typeIndice:entiteNom;typePos:AMORCE	9916.000
3.42	0.000	0.034	128. V131-title->entiteNom.classe:lieu;sousClasse:indetermine	9916.000
3.29	0.001	0.033	90. V093-exprTemp.nature:deictique;sitTps:posteriorite	9916.000
3.14	0.001	0.031	6. V005-tpsVbx.temps:passeComp	9916.000
2.96	0.002	0.030	49. V048-position.typeIndice:ptVue;typePos:IC	9916.000
2.95	0.002	0.030	59. V058-argum.relation:opposition	9916.000
2.85	0.002	0.029	89. V092-zone.rubriqueName:FauneFlore	9916.000
2.75	0.003	0.028	134. V137-title->entiteNom.classe:lieu;sousClasse:pays	9916.000
2.63	0.004	0.026	77. V080-position.typeIndice:ptVue;typePos:AMORCE	9916.000

2.49	0.006	0.025	15.	V014-argum.relation:identite	9916.000
2.48	0.007	0.025	47.	V046-argum.relation:liste	9916.000
2.45	0.007	0.025	69.	V071-exprTemp.nature:duree;sitTps:coincidence	9916.000
2.42	0.008	0.024	60.	V062-entiteNom.classe:lieu;sousClasse:ville	9916.000
2.35	0.009	0.024	107.	V110-zone.rubriqueName:	9916.000
2.33	0.010	0.023	39.	V038-ptVue.type:recence	9916.000
2.27	0.011	0.023	36.	V035-descPhrasePosition.position:seuleInParagraphe	9916.000
2.20	0.014	0.022	96.	V099-zone.rubriqueName:ScTechn	9916.000
2.17	0.015	0.022	45.	V044-position.typeIndice:exprTemp;typePos:END	9916.000
2.16	0.015	0.022	16.	V015-argum.relation:correction	9916.000
2.14	0.016	0.021	62.	V064-exprTemp.nature:duree;sitTps:indetermine	9916.000
2.09	0.018	0.021	44.	V043-exprTemp.nature:inachevee;sitTps:anteriorite	9916.000
2.07	0.019	0.021	5.	V004-exprTemp.nature:iteration;sitTps:indetermine	9916.000
1.99	0.023	0.020	71.	V073-argum.relation:incertitude	9916.000
1.97	0.025	0.020	95.	V098-ptVue.type:importance	9916.000
1.93	0.027	0.019	144.	V147-title->argum.relation:correction	9916.000
1.88	0.030	0.019	127.	V130-title->ptVue.type:recence	9916.000
1.87	0.031	0.019	145.	V148-title->ptVue.type:importance	9916.000
1.84	0.033	0.018	57.	V056-entiteNom.classe:mesure;sousClasse:indetermine	9916.000
1.83	0.033	0.018	53.	V052-position.typeIndice:exprTemp;typePos:IC	9916.000
1.83	0.034	0.018	120.	V123-title->exprTemp.nature:deictique;sitTps:coincidence	9916.000
1.82	0.034	0.018	80.	V083-entiteNom.classe:lieu;sousClasse:riviere	9916.000
1.80	0.036	0.018	9.	V008-argum.relation:precision	9916.000
1.72	0.042	0.017	106.	V109-entiteNom.classe:moins	9916.000
1.64	0.050	0.017	63.	V065-ptVue.type:jugement	9916.000
1.49	0.068	0.015	124.	V127-title->tpsVbx.temps:conditionnel	9916.000
1.47	0.071	0.015	25.	V024-tpsVbx.temps:plusQuePft	9916.000
1.42	0.078	0.014	109.	V112-descPhraseType.type:exclamation	9916.000
1.35	0.088	0.014	88.	V091-premierParag.position:debutZone	9916.000
1.29	0.098	0.013	54.	V053-premierParag.position:debutDivision	9916.000
1.16	0.123	0.012	84.	V087-periVbs.accomplissement:fin	9916.000
1.16	0.123	0.012	146.	V149-title->exprTemp.nature:inachevee;sitTps:anteriorite	9916.000
0.96	0.169	0.010	76.	V079-tpsVbx.temps:condPasse	9916.000
0.92	0.180	0.009	136.	V139-title->entiteNom.classe:lieu;sousClasse:ville	9916.000
0.84	0.199	0.008	33.	V032-position.typeIndice:argum;typePos:IC	9916.000
0.83	0.203	0.008	131.	V134-title->ptVue.type:prevision	9916.000
0.77	0.221	0.008	68.	V070-exprTemp.nature:anaphorique;sitTps:posteriorite	9916.000
0.72	0.236	0.007	14.	V013-tpsVbx.temps:futur	9916.000
0.69	0.245	0.007	137.	V140-title->tpsVbx.temps:imparfait	9916.000
0.54	0.294	0.005	52.	V051-argum.relation:concession	9916.000
0.50	0.310	0.005	19.	V018-argum.relation:explication	9916.000
0.49	0.313	0.005	37.	V036-descPhraseType.type:interrogation	9916.000
0.41	0.339	0.004	51.	V050-ptVue.type:restriction	9916.000
0.40	0.345	0.004	99.	V102-exprTemp.nature:inachevee;sitTps:indetermine	9916.000
0.26	0.397	0.003	26.	V025-tpsVbx.temps:imparfait	9916.000
0.26	0.398	0.003	83.	V086-entiteNom.classe:mesure;sousClasse:fixe	9916.000
0.20	0.421	0.002	58.	V057-tpsVbx.temps:futAnt	9916.000
0.19	0.425	0.002	82.	V085-zone.rubriqueName:Soc	9916.000
0.16	0.435	0.002	86.	V089-entiteNom.classe:lieu;sousClasse:adresse	9916.000
0.01	0.495	0.000	138.	V141-title->entiteNom.classe:lieu;sousClasse:riviere	9916.000
-0.01	0.496	0.000	2.	V001-descPhrasePosition.position:debutParagraphe	9916.000
-0.25	0.402	-0.002	122.	V125-title->exprTemp.nature:ponctuel;sitTps:coincidence	9916.000
-0.29	0.385	-0.003	140.	V143-title->tpsVbx.temps:passeSimple	9916.000
-0.36	0.359	-0.004	112.	V115-exprTemp.nature:ponctuel	9916.000
-0.36	0.359	-0.004	111.	V114-ptVue.type:enonciatif	9916.000
-0.36	0.359	-0.004	110.	V113-exprTemp.nature:inachevee;sitTps:posteriorite	9916.000
-0.48	0.316	-0.005	40.	V039-argum.relation:justification	9916.000
-0.49	0.312	-0.005	142.	V145-title->tpsVbx.temps:futur	9916.000
-0.51	0.305	-0.005	94.	V097-exprTemp.nature:deictique;sitTps:indetermine	9916.000
-0.51	0.305	-0.005	141.	V144-title->exprTemp.nature:inachevee;sitTps:trespasse	9916.000
-0.51	0.305	-0.005	102.	V105-exprTemp.nature:duree;sitTps:trespasse	9916.000
-0.56	0.289	-0.006	67.	V069-exprTemp.nature:anaphorique;sitTps:coincidence	9916.000

-0.58	0.280	-0.006	34. V033-argum.relation:illustration	9916.000
-0.65	0.258	-0.007	97. V100-premierParag.position:debutZoneSeul	9916.000
-0.66	0.255	-0.007	114. V117-title->niveau:1	9916.000
-0.67	0.252	-0.007	32. V031-ptVue.type:distance	9916.000
-0.72	0.236	-0.007	104. V107-exprTemp.nature:anaphorique;sitTps:anteriorite	9916.000
-0.79	0.216	-0.008	139. V142-title->position.typeIndice:entiteNom;typePos:AMORCE	9916.000
-0.79	0.214	-0.008	126. V129-title->niveau:3	9916.000
-0.87	0.191	-0.009	42. V041-position.typeIndice:ptVue;typePos:END	9916.000
-0.94	0.174	-0.009	119. V122-title->argum.relation:temporelle	9916.000
-0.95	0.172	-0.010	55. V054-exprTemp.nature:ponctuel;sitTps:indetermine	9916.000
-0.95	0.170	-0.010	101. V104-position.typeIndice:argum;typePos:END	9916.000
-0.99	0.161	-0.010	81. V084-ptVue.type:definition	9916.000
-1.02	0.154	-0.010	130. V133-title->niveau:5	9916.000
-1.09	0.138	-0.011	61. V063-exprTemp.nature:anaphorique;sitTps:indetermine	9916.000
-1.19	0.118	-0.012	132. V135-title->entiteNom.classe:sigle	9916.000
-1.39	0.083	-0.014	18. V017-argum.relation:consecution	9916.000
-1.39	0.082	-0.014	3. V002-descPhraseType.type:assertion	9916.000
-1.45	0.073	-0.015	103. V106-position.typeIndice:exprTemp;typePos:AMORCE	9916.000
-1.62	0.053	-0.016	38. V037-periVbs.accomplissement:debut	9916.000
-1.65	0.049	-0.017	123. V126-title->ptVue.type:distance	9916.000
-1.80	0.036	-0.018	70. V072-tpsVbx.temps:passeAnt	9916.000
-1.81	0.035	-0.018	50. V049-exprTemp.nature:inachevee;sitTps:trespasse	9916.000
-1.84	0.033	-0.018	92. V095-entiteNom.classe:lieu;sousClasse:pointCardinal	9916.000
-2.11	0.017	-0.021	31. V030-exprTemp.nature:deictique;sitTps:trespasse	9916.000
-2.16	0.015	-0.022	118. V121-title->tpsVbx.temps:passeComp	9916.000
-2.37	0.009	-0.024	7. V006-tpsVbx.temps:present	9916.000
-2.41	0.008	-0.024	41. V040-tpsVbx.temps:passeSimple	9916.000
-2.53	0.006	-0.025	8. V007-zone.rubriqueName:Eco	9916.000
-3.18	0.001	-0.032	115. V118-title->tpsVbx.temps:present	9916.000
-3.45	0.000	-0.035	10. V009-argum.relation:temporelle	9916.000
-4.09	0.000	-0.041	64. V066-premierParag.position:debutDivisionSeul	9916.000
-4.39	0.000	-0.044	116. V119-title->niveau:2	9916.000
-4.54	0.000	-0.046	121. V124-title->exprTemp.nature:ponctuel;sitTps:anteriorite	9916.000
-4.75	0.000	-0.048	13. V012-exprTemp.nature:ponctuel;sitTps:anteriorite	9916.000
-4.95	0.000	-0.050	135. V138-title->exprTemp.nature:ponctuel;sitTps:trespasse	9916.000
-5.48	0.000	-0.055	125. V128-title->entiteNom.classe:personne	9916.000
-5.95	0.000	-0.060	20. V019-entiteNom.classe:personne	9916.000
-6.18	0.000	-0.062	24. V023-exprTemp.nature:ponctuel;sitTps:trespasse	9916.000
-6.32	0.000	-0.063	66. V068-zone.rubriqueName:Hist	9916.000
-6.42	0.000	-0.064	108. V111-zone.rubriqueName:Sport	9916.000
-6.68	0.000	-0.067	113. V116-zone.rubriqueName:ArtLitt	9916.000
-10.30	0.000	-0.103	98. V101-zone.rubriqueName:Med	9916.000

C.2.2 Statistiques sommaires des variables continues

EFFECTIF TOTAL 9916
POIDS TOTAL 9916.00

NUM. IDEN - LIBELLE	EFFECTIF	POIDS	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
1. Obs - Obs	9916	9916.00	0.20	0.57	0.00	4.00
2. V001 - V001-descPhrasePosit	9916	9916.00	0.16	0.37	0.00	1.00
3. V002 - V002-descPhraseType.	9916	9916.00	0.97	0.17	0.00	1.00
5. V004 - V004-exprTemp.nature	9916	9916.00	0.02	0.14	0.00	2.00
6. V005 - V005-tpsVbx.temps:pa	9916	9916.00	0.27	0.51	0.00	5.00
7. V006 - V006-tpsVbx.temps:pr	9916	9916.00	1.12	0.98	0.00	11.00
8. V007 - V007-zone.rubriqueNa	9916	9916.00	0.07	0.25	0.00	1.00
9. V008 - V008-argum.relation:	9916	9916.00	0.02	0.13	0.00	2.00
10. V009 - V009-argum.relation:	9916	9916.00	0.04	0.21	0.00	3.00
11. V010 - V010-entiteNom.class	9916	9916.00	0.15	0.44	0.00	6.00
12. V011 - V011-exprTemp.nature	9916	9916.00	0.00	0.06	0.00	1.00
13. V012 - V012-exprTemp.nature	9916	9916.00	0.09	0.37	0.00	9.00
14. V013 - V013-tpsVbx.temps:fu	9916	9916.00	0.03	0.20	0.00	4.00
15. V014 - V014-argum.relation:	9916	9916.00	0.03	0.17	0.00	2.00
16. V015 - V015-argum.relation:	9916	9916.00	0.06	0.23	0.00	2.00
17. V016 - V016-descPhrasePosit	9916	9916.00	0.16	0.36	0.00	1.00
18. V017 - V017-argum.relation:	9916	9916.00	0.04	0.20	0.00	2.00
19. V018 - V018-argum.relation:	9916	9916.00	0.06	0.24	0.00	3.00
20. V019 - V019-entiteNom.class	9916	9916.00	0.18	0.67	0.00	15.00
21. V020 - V020-entiteNom.class	9916	9916.00	0.10	0.44	0.00	13.00
22. V021 - V021-entiteNom.class	9916	9916.00	0.14	0.51	0.00	15.00
23. V022 - V022-position.typeIn	9916	9916.00	0.06	0.23	0.00	1.00
24. V023 - V023-exprTemp.nature	9916	9916.00	0.12	0.52	0.00	18.00
25. V024 - V024-tpsVbx.temps:pl	9916	9916.00	0.01	0.10	0.00	2.00
26. V025 - V025-tpsVbx.temps:im	9916	9916.00	0.04	0.24	0.00	4.00
27. V026 - V026-entiteNom.class	9916	9916.00	0.04	0.20	0.00	2.00
28. V027 - V027-tpsVbx.temps:co	9916	9916.00	0.02	0.15	0.00	2.00
29. V028 - V028-exprTemp.nature	9916	9916.00	0.04	0.23	0.00	4.00
30. V029 - V029-ptVue.type:prev	9916	9916.00	0.02	0.16	0.00	2.00
31. V030 - V030-exprTemp.nature	9916	9916.00	0.00	0.07	0.00	2.00
32. V031 - V031-ptVue.type:dist	9916	9916.00	0.00	0.06	0.00	2.00
33. V032 - V032-position.typeIn	9916	9916.00	0.02	0.16	0.00	2.00
34. V033 - V033-argum.relation:	9916	9916.00	0.01	0.10	0.00	2.00
35. V034 - V034-exprTemp.nature	9916	9916.00	0.07	0.27	0.00	3.00
36. V035 - V035-descPhrasePosit	9916	9916.00	0.07	0.25	0.00	1.00
37. V036 - V036-descPhraseType.	9916	9916.00	0.00	0.05	0.00	1.00
38. V037 - V037-periVbs.accompl	9916	9916.00	0.01	0.08	0.00	1.00
39. V038 - V038-ptVue.type:rece	9916	9916.00	0.04	0.21	0.00	3.00
40. V039 - V039-argum.relation:	9916	9916.00	0.01	0.11	0.00	1.00
41. V040 - V040-tpsVbx.temps:pa	9916	9916.00	0.04	0.24	0.00	6.00
42. V041 - V041-position.typeIn	9916	9916.00	0.01	0.09	0.00	1.00
43. V042 - V042-dernierParag.po	9916	9916.00	0.17	0.38	0.00	1.00
44. V043 - V043-exprTemp.nature	9916	9916.00	0.02	0.14	0.00	5.00
45. V044 - V044-position.typeIn	9916	9916.00	0.05	0.22	0.00	1.00
46. V045 - V045-entiteNom.class	9916	9916.00	0.17	0.62	0.00	12.00
47. V046 - V046-argum.relation:	9916	9916.00	0.02	0.14	0.00	3.00
48. V047 - V047-position.typeIn	9916	9916.00	0.13	0.34	0.00	1.00
49. V048 - V048-position.typeIn	9916	9916.00	0.01	0.09	0.00	1.00
50. V049 - V049-exprTemp.nature	9916	9916.00	0.01	0.11	0.00	2.00
51. V050 - V050-ptVue.type:rest	9916	9916.00	0.01	0.09	0.00	4.00
52. V051 - V051-argum.relation:	9916	9916.00	0.00	0.03	0.00	1.00
53. V052 - V052-position.typeIn	9916	9916.00	0.06	0.24	0.00	1.00
54. V053 - V053-premierParag.po	9916	9916.00	0.03	0.16	0.00	1.00
55. V054 - V054-exprTemp.nature	9916	9916.00	0.01	0.08	0.00	2.00

56. V055 - V055-dernierParag.po	9916	9916.00	0.06	0.23	0.00	1.00
57. V056 - V056-entiteNom.class	9916	9916.00	0.12	1.22	0.00	75.00
58. V057 - V057-tpsVbx.temps:fu	9916	9916.00	0.00	0.06	0.00	1.00
59. V058 - V058-argum.relation:	9916	9916.00	0.03	0.16	0.00	2.00
60. V062 - V062-entiteNom.class	9916	9916.00	0.10	0.59	0.00	27.00
61. V063 - V063-exprTemp.nature	9916	9916.00	0.01	0.13	0.00	4.00
62. V064 - V064-exprTemp.nature	9916	9916.00	0.01	0.11	0.00	2.00
63. V065 - V065-ptVue.type:juge	9916	9916.00	0.00	0.06	0.00	1.00
64. V066 - V066-premierParag.po	9916	9916.00	0.06	0.24	0.00	1.00
65. V067 - V067-zone.rubriqueNa	9916	9916.00	0.00	0.04	0.00	1.00
66. V068 - V068-zone.rubriqueNa	9916	9916.00	0.15	0.36	0.00	1.00
67. V069 - V069-exprTemp.nature	9916	9916.00	0.00	0.04	0.00	1.00
68. V070 - V070-exprTemp.nature	9916	9916.00	0.00	0.06	0.00	2.00
69. V071 - V071-exprTemp.nature	9916	9916.00	0.00	0.05	0.00	1.00
70. V072 - V072-tpsVbx.temps:pa	9916	9916.00	0.00	0.07	0.00	3.00
71. V073 - V073-argum.relation:	9916	9916.00	0.00	0.01	0.00	1.00
72. V074 - V074-argum.relation:	9916	9916.00	0.00	0.03	0.00	1.00
73. V076 - V076-position.typeIn	9916	9916.00	0.03	0.19	0.00	6.00
74. V077 - V077-entiteNom.class	9916	9916.00	0.06	0.31	0.00	6.00
75. V078 - V078-entiteNom.class	9916	9916.00	0.02	0.15	0.00	3.00
76. V079 - V079-tpsVbx.temps:co	9916	9916.00	0.00	0.06	0.00	2.00
77. V080 - V080-position.typeIn	9916	9916.00	0.00	0.04	0.00	1.00
78. V081 - V081-ptVue.type:sour	9916	9916.00	0.00	0.05	0.00	1.00
79. V082 - V082-exprTemp.nature	9916	9916.00	0.01	0.08	0.00	1.00
80. V083 - V083-entiteNom.class	9916	9916.00	0.02	0.19	0.00	4.00
81. V084 - V084-ptVue.type:defi	9916	9916.00	0.00	0.05	0.00	1.00
82. V085 - V085-zone.rubriqueNa	9916	9916.00	0.12	0.33	0.00	1.00
83. V086 - V086-entiteNom.class	9916	9916.00	0.06	0.31	0.00	6.00
84. V087 - V087-periVbs.accompl	9916	9916.00	0.00	0.04	0.00	1.00
85. V088 - V088-exprTemp.nature	9916	9916.00	0.00	0.05	0.00	2.00
86. V089 - V089-entiteNom.class	9916	9916.00	0.00	0.02	0.00	1.00
87. V090 - V090-periVbs.accompl	9916	9916.00	0.00	0.02	0.00	1.00
88. V091 - V091-premierParag.po	9916	9916.00	0.01	0.12	0.00	1.00
89. V092 - V092-zone.rubriqueNa	9916	9916.00	0.03	0.16	0.00	1.00
90. V093 - V093-exprTemp.nature	9916	9916.00	0.00	0.06	0.00	1.00
91. V094 - V094-zone.rubriqueNa	9916	9916.00	0.18	0.39	0.00	1.00
92. V095 - V095-entiteNom.class	9916	9916.00	0.02	0.18	0.00	7.00
93. V096 - V096-entiteNom.class	9916	9916.00	0.00	0.05	0.00	1.00
94. V097 - V097-exprTemp.nature	9916	9916.00	0.00	0.01	0.00	1.00
95. V098 - V098-ptVue.type:impo	9916	9916.00	0.00	0.05	0.00	1.00
96. V099 - V099-zone.rubriqueNa	9916	9916.00	0.13	0.33	0.00	1.00
97. V100 - V100-premierParag.po	9916	9916.00	0.03	0.18	0.00	1.00
98. V101 - V101-zone.rubriqueNa	9916	9916.00	0.17	0.38	0.00	1.00
99. V102 - V102-exprTemp.nature	9916	9916.00	0.00	0.02	0.00	1.00
101. V104 - V104-position.typeIn	9916	9916.00	0.00	0.03	0.00	1.00
102. V105 - V105-exprTemp.nature	9916	9916.00	0.00	0.01	0.00	1.00
103. V106 - V106-position.typeIn	9916	9916.00	0.00	0.06	0.00	1.00
104. V107 - V107-exprTemp.nature	9916	9916.00	0.00	0.02	0.00	1.00
106. V109 - V109-entiteNom.class	9916	9916.00	0.00	0.03	0.00	1.00
107. V110 - V110-zone.rubriqueNa	9916	9916.00	0.06	0.23	0.00	1.00
108. V111 - V111-zone.rubriqueNa	9916	9916.00	0.05	0.22	0.00	1.00
109. V112 - V112-descPhraseType.	9916	9916.00	0.00	0.02	0.00	1.00
110. V113 - V113-exprTemp.nature	9916	9916.00	0.00	0.01	0.00	1.00
111. V114 - V114-ptVue.type:enon	9916	9916.00	0.00	0.01	0.00	1.00
112. V115 - V115-exprTemp.nature	9916	9916.00	0.00	0.01	0.00	1.00
113. V116 - V116-zone.rubriqueNa	9916	9916.00	0.03	0.18	0.00	1.00
114. V117 - V117-title->niveau:1	9916	9916.00	0.25	0.43	0.00	1.00
115. V118 - V118-title->tpsVbx.t	9916	9916.00	0.03	0.18	0.00	1.00
116. V119 - V119-title->niveau:2	9916	9916.00	0.54	0.50	0.00	1.00
117. V120 - V120-title->entiteNo	9916	9916.00	0.06	0.25	0.00	1.00
118. V121 - V121-title->tpsVbx.t	9916	9916.00	0.00	0.06	0.00	1.00
119. V122 - V122-title->argum.re	9916	9916.00	0.00	0.04	0.00	1.00

120. V123 - V123-title->exprTemp	9916	9916.00	0.00	0.05	0.00	1.00
121. V124 - V124-title->exprTemp	9916	9916.00	0.02	0.15	0.00	2.00
122. V125 - V125-title->exprTemp	9916	9916.00	0.01	0.09	0.00	1.00
123. V126 - V126-title->ptVue.ty	9916	9916.00	0.00	0.05	0.00	1.00
124. V127 - V127-title->tpsVbx.t	9916	9916.00	0.00	0.04	0.00	1.00
125. V128 - V128-title->entiteNo	9916	9916.00	0.05	0.24	0.00	3.00
126. V129 - V129-title->niveau:3	9916	9916.00	0.12	0.32	0.00	1.00
127. V130 - V130-title->ptVue.ty	9916	9916.00	0.01	0.11	0.00	1.00
128. V131 - V131-title->entiteNo	9916	9916.00	0.02	0.13	0.00	1.00
129. V132 - V132-title->niveau:4	9916	9916.00	0.01	0.08	0.00	1.00
130. V133 - V133-title->niveau:5	9916	9916.00	0.00	0.03	0.00	1.00
131. V134 - V134-title->ptVue.ty	9916	9916.00	0.01	0.09	0.00	1.00
132. V135 - V135-title->entiteNo	9916	9916.00	0.02	0.14	0.00	2.00
133. V136 - V136-title->niveau:0	9916	9916.00	0.07	0.25	0.00	1.00
134. V137 - V137-title->entiteNo	9916	9916.00	0.04	0.19	0.00	1.00
135. V138 - V138-title->exprTemp	9916	9916.00	0.03	0.16	0.00	1.00
136. V139 - V139-title->entiteNo	9916	9916.00	0.02	0.15	0.00	2.00
137. V140 - V140-title->tpsVbx.t	9916	9916.00	0.00	0.03	0.00	1.00
138. V141 - V141-title->entiteNo	9916	9916.00	0.00	0.07	0.00	2.00
139. V142 - V142-title->position	9916	9916.00	0.00	0.06	0.00	1.00
140. V143 - V143-title->tpsVbx.t	9916	9916.00	0.01	0.07	0.00	1.00
141. V144 - V144-title->exprTemp	9916	9916.00	0.00	0.01	0.00	1.00
142. V145 - V145-title->tpsVbx.t	9916	9916.00	0.00	0.05	0.00	1.00
143. V146 - V146-title->exprTemp	9916	9916.00	0.00	0.03	0.00	1.00
144. V147 - V147-title->argum.re	9916	9916.00	0.00	0.02	0.00	1.00
145. V148 - V148-title->ptVue.ty	9916	9916.00	0.00	0.03	0.00	1.00
146. V149 - V149-title->exprTemp	9916	9916.00	0.00	0.04	0.00	1.00

C.3 ACP

C.3.1 Histogramme des 146 premières valeurs propres

```
VALEURS PROPRES
APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 146.0000
                                     SOMME DES VALEURS PROPRES .... 146.0000
HISTOGRAMME DES146 PREMIERES VALEURS PROPRES
```

NUMERO	VALEUR PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	3.3371	2.29	2.29	**********
2	2.9325	2.01	4.29	*********
3	2.3904	1.64	5.93	*******
4	2.2556	1.54	7.48	*******
5	2.0876	1.43	8.91	******
6	2.0218	1.38	10.29	******
7	1.8460	1.26	11.56	*****
8	1.7654	1.21	12.76	*****
9	1.6613	1.14	13.90	*****
10	1.6152	1.11	15.01	*****
11	1.5853	1.09	16.09	*****
12	1.5474	1.06	17.15	*****
13	1.5054	1.03	18.19	*****
14	1.4691	1.01	19.19	*****
15	1.4514	0.99	20.19	*****
16	1.4136	0.97	21.15	*****
17	1.3805	0.95	22.10	*****
18	1.3633	0.93	23.03	*****
19	1.3318	0.91	23.95	*****
20	1.3117	0.90	24.84	*****
21	1.3080	0.90	25.74	*****
22	1.2954	0.89	26.63	*****
23	1.2737	0.87	27.50	*****
24	1.2626	0.86	28.36	*****
25	1.2430	0.85	29.22	*****
26	1.2339	0.85	30.06	*****

27	1.2262	0.84	30.90	******************************
28	1.2196	0.84	31.74	******************************
29	1.2085	0.83	32.56	*****************************
30	1.1979	0.82	33.38	*****************************
31	1.1858	0.81	34.20	*****************************
32	1.1755	0.81	35.00	*****************************
33	1.1643	0.80	35.80	****************************
34	1.1628	0.80	36.60	****************************
35	1.1513	0.79	37.38	****************************
36	1.1365	0.78	38.16	****************************
37	1.1304	0.77	38.94	****************************
38	1.1283	0.77	39.71	****************************
39	1.1244	0.77	40.48	****************************
40	1.1128	0.76	41.24	****************************
41	1.1054	0.76	42.00	****************************
42	1.0960	0.75	42.75	****************************
43	1.0823	0.74	43.49	**************************
44	1.0803	0.74	44.23	**************************
45	1.0793	0.74	44.97	**************************
46	1.0731	0.74	45.70	**************************
47	1.0693	0.73	46.44	**************************
48	1.0646	0.73	47.17	**************************
49	1.0581	0.72	47.89	**************************
50	1.0517	0.72	48.61	**************************
51	1.0476	0.72	49.33	**************************
52	1.0402	0.71	50.04	************************
53	1.0380	0.71	50.75	************************
54	1.0363	0.71	51.46	************************
55	1.0332	0.71	52.17	************************
56	1.0303	0.71	52.88	************************
57	1.0243	0.70	53.58	************************
58	1.0223	0.70	54.28	************************
59	1.0204	0.70	54.98	************************
60	1.0143	0.69	55.67	************************
61	1.0120	0.69	56.36	************************
62	1.0091	0.69	57.06	************************
63	1.0067	0.69	57.75	************************
64	1.0053	0.69	58.43	************************
65	1.0007	0.69	59.12	**********************
66	0.9975	0.68	59.80	**********************
67	0.9960	0.68	60.48	**********************
68	0.9910	0.68	61.16	**********************
69	0.9887	0.68	61.84	**********************
70	0.9858	0.68	62.52	**********************
71	0.9824	0.67	63.19	**********************
72	0.9781	0.67	63.86	**********************
73	0.9755	0.67	64.53	**********************
74	0.9691	0.66	65.19	**********************
75	0.9682	0.66	65.85	**********************
76	0.9610	0.66	66.51	**********************
77	0.9566	0.66	67.17	**********************
78	0.9504	0.65	67.82	**********************
79	0.9468	0.65	68.47	**********************
80	0.9448	0.65	69.11	**********************
81	0.9415	0.64	69.76	**********************
82	0.9334	0.64	70.40	**********************
83	0.9298	0.64	71.03	**********************
84	0.9250	0.63	71.67	**********************
85	0.9221	0.63	72.30	**********************
86	0.9169	0.63	72.93	**********************
87	0.9116	0.62	73.55	**********************
88	0.9043	0.62	74.17	**********************
89	0.8993	0.62	74.79	**********************
90	0.8988	0.62	75.40	**********************
91	0.8912	0.61	76.01	**********************
92	0.8839	0.61	76.62	**********************
93	0.8806	0.60	77.22	**********************
94	0.8769	0.60	77.82	**********************
95	0.8690	0.60	78.42	**********************
96	0.8646	0.59	79.01	**********************
97	0.8623	0.59	79.60	**********************
98	0.8558	0.59	80.19	**********************
99	0.8499	0.58	80.77	**********************
100	0.8469	0.58	81.35	**********************
101	0.8411	0.58	81.92	**********************
102	0.8367	0.57	82.50	**********************
103	0.8296	0.57	83.07	**********************
104	0.8266	0.57	83.63	**********************
105	0.8149	0.56	84.19	**********************
106	0.8081	0.55	84.74	**********************
107	0.8015	0.55	85.29	**********************
108	0.7885	0.54	85.83	********************
109	0.7819	0.54	86.37	********************
110	0.7761	0.53	86.90	********************

111	0.7624	0.52	87.42	******************
112	0.7595	0.52	87.94	******************
113	0.7546	0.52	88.46	******************
114	0.7423	0.51	88.97	*****************
115	0.7343	0.50	89.47	*****************
116	0.7312	0.50	89.97	*****************
117	0.7159	0.49	90.46	*****************
118	0.7034	0.48	90.94	****************
119	0.6990	0.48	91.42	****************
120	0.6832	0.47	91.89	****************
121	0.6776	0.46	92.35	****************
122	0.6735	0.46	92.82	****************
123	0.6642	0.45	93.27	***************
124	0.6569	0.45	93.72	***************
125	0.6435	0.44	94.16	***************
126	0.6381	0.44	94.60	***************
127	0.6236	0.43	95.03	**************
128	0.6165	0.42	95.45	**************
129	0.5867	0.40	95.85	**************
130	0.5686	0.39	96.24	*************
131	0.5543	0.38	96.62	*************
132	0.5391	0.37	96.99	************
133	0.5038	0.35	97.33	************
134	0.4962	0.34	97.67	************
135	0.4742	0.32	98.00	**********
136	0.4549	0.31	98.31	**********
137	0.4430	0.30	98.61	**********
138	0.4197	0.29	98.90	**********
139	0.3981	0.27	99.17	*********
140	0.3804	0.26	99.43	*********
141	0.3552	0.24	99.68	********
142	0.3056	0.21	99.89	********
143	0.1450	0.10	99.99	****
144	0.0215	0.01	100.00	*
145	0.0000	0.00	100.00	*
146	0.0000	0.00	100.00	*

C.3.2 Recherche de paliers (différences troisièmes)

RECHERCHE DE PALIERS (DIFFERENCES TROISIEMES)

PALIER ENTRE	VALEUR DU PALIER	
2 -- 3	-440.57	**
6 -- 7	-212.35	*************************
132 --133	-42.25	*****
8 -- 9	-41.81	*****
15 -- 16	-38.65	*****
16 -- 17	-30.27	****
128 --129	-30.24	****
19 -- 20	-25.41	***
9 -- 10	-24.41	***
34 -- 35	-19.62	***
119 --120	-19.55	***
22 -- 23	-19.08	***
137 --138	-18.86	***
126 --127	-17.23	***
116 --117	-17.05	***
141 --142	-16.99	***
110 --111	-12.80	**
104 --105	-12.53	**
75 -- 76	-11.82	**
90 -- 91	-11.63	**
140 --141	-11.41	**
42 -- 43	-10.76	**
24 -- 25	-9.21	**
124 --125	-8.05	*
107 --108	-6.79	*
122 --123	-6.63	*
97 -- 98	-6.24	*
45 -- 46	-6.13	*
39 -- 40	-6.04	*
28 -- 29	-5.81	*
98 -- 99	-5.72	*
36 -- 37	-5.65	*
81 -- 82	-5.60	*
113 --114	-5.42	*
51 -- 52	-5.34	*
73 -- 74	-5.17	*
67 -- 68	-5.07	*
129 --130	-4.86	*
35 -- 36	-4.82	*
91 -- 92	-4.30	*
85 -- 86	-4.29	*
59 -- 60	-4.23	*
64 -- 65	-4.17	*
102 --103	-4.02	*
56 -- 57	-3.88	*
94 -- 95	-3.76	*
78 -- 79	-2.81	*
30 -- 31	-2.62	*
54 -- 55	-2.02	*
87 -- 88	-1.90	*
77 -- 78	-0.96	*
48 -- 49	-0.94	*
71 -- 72	-0.25	*
25 -- 26	-0.11	*
70 -- 71	-0.05	*

C.3.3 Recherche de paliers (différences secondes)

RECHERCHE DE PALIERS ENTRE (DIFFERENCES SECONDES)

PALIER ENTRE	VALEUR DU PALIER	
2 -- 3	407.29	**
4 -- 5	102.28	**************
6 -- 7	95.34	*************
8 -- 9	58.07	********
132 --133	27.75	****
13 -- 14	18.55	***
19 -- 20	16.42	***
9 -- 10	16.27	***
16 -- 17	15.89	***
42 -- 43	11.74	**
128 --129	11.71	**
18 -- 19	11.46	**
110 --111	10.83	**
22 -- 23	10.57	**
24 -- 25	10.52	**
119 --120	10.20	**
32 -- 33	9.65	**
35 -- 36	8.73	**
117 --118	8.10	**
124 --125	8.03	**
126 --127	7.49	*
135 --136	7.40	*
107 --108	6.36	*
12 -- 13	5.75	*
73 -- 74	5.44	*
51 -- 52	5.10	*
114 --115	4.88	*
104 --105	4.77	*
15 -- 16	4.74	*
88 -- 89	4.55	*
81 -- 82	4.47	*
113 --114	4.26	*
39 -- 40	4.12	*
102 --103	3.96	*
56 -- 57	3.94	*
138 --139	3.93	*
91 -- 92	3.92	*
36 -- 37	3.91	*
129 --130	3.86	*
59 -- 60	3.69	*
94 -- 95	3.43	*
75 -- 76	2.96	*
98 -- 99	2.93	*
116 --117	2.83	*
134 --135	2.76	*
77 -- 78	2.56	*
67 -- 68	2.53	*
45 -- 46	2.38	*
87 -- 88	2.18	*
49 -- 50	2.18	*
95 -- 96	2.11	*
122 --123	1.96	*
83 -- 84	1.92	*
30 -- 31	1.76	*
137 --138	1.70	*
65 -- 66	1.68	*
78 -- 79	1.60	*
71 -- 72	1.52	*
120 --121	1.49	*
64 -- 65	1.41	*
100 --101	1.39	*
25 -- 26	1.32	*
26 -- 27	1.21	*
43 -- 44	0.99	*
62 -- 63	0.95	*
108 --109	0.86	*
28 -- 29	0.56	*
52 -- 53	0.55	*
61 -- 62	0.54	*
97 -- 98	0.49	*
90 -- 91	0.35	*
105 --106	0.32	*
54 -- 55	0.30	*
48 -- 49	0.21	*
57 -- 58	0.06	*

```
+--------------+--------------+-----------------------------------------------------+
```

C.3.4 Intervalles Laplaciens d'Anderson

```
INTERVALLES LAPLACIENS D'ANDERSON
INTERVALLES AU SEUIL 0.95
+--------+-----------------------------------------------------+
| NUMERO | BORNE INFERIEURE     VALEUR PROPRE    BORNE SUPERIEURE |
+--------+-----------------------------------------------------+
|    1   |      3.2442             3.3371             3.4300      |
|    2   |      2.8508             2.9325             3.0142      |
|    3   |      2.3239             2.3904             2.4570      |
|    4   |      2.1928             2.2556             2.3185      |
|    5   |      2.0294             2.0876             2.1457      |
+--------+-----------------------------------------------------+
ETENDUE ET POSITION RELATIVE DES INTERVALLES
1 . . . . . .*-------+--------*
2 . . . . . .*------+------*. . . . . .
3 . . . . . *-----+------*. . . . . .
4 . . . . . *-----+-----* . . . . . .
5 *----+----* . . . . . .
```

C.4 DEFAC : description des axes factoriels

DESCRIPTION DES AXES FACTORIELS
DESCRIPTION DU FACTEUR 1
PAR LES INDIVIDUS ACTIFS

COORD.	POIDS	IDENTIFICATEUR DE L'INDIVIDU	NUMERO
-14.35	1.00	1238262901418	1
-10.75	1.00	1238262896783	2
-10.08	1.00	1238262896785	3
-10.03	1.00	1238262898095	4
-10.00	1.00	1238262902106	5
-9.97	1.00	1238262900190	6
-9.79	1.00	1238262900828	7
-9.76	1.00	1238262902534	8
-9.75	1.00	1238262902405	9
-9.69	1.00	1238262901025	10
-9.68	1.00	1238262901296	11
-9.58	1.00	1238262899015	12
-9.58	1.00	1238262899014	13
-9.58	1.00	1238262900027	14
-9.58	1.00	1238262900028	15
-9.57	1.00	1238262902987	16
-9.57	1.00	1238262902404	17
-9.50	1.00	1238262900490	18
-9.50	1.00	1238262900532	19
-9.33	1.00	1238262903202	20
		Z O N E C E N T R A L E	
3.44	1.00	1237456509806	9891
3.45	1.00	1237456511045	9892
3.47	1.00	1237456509793	9893
3.53	1.00	1237456509805	9894
3.54	1.00	1237456509787	9895
3.54	1.00	1237456511547	9896
3.55	1.00	1237456509810	9897
3.56	1.00	1237456510304	9898
3.58	1.00	1238262897702	9899
3.63	1.00	1237456511042	9900
3.69	1.00	1237456510228	9901
3.69	1.00	1237456511004	9902
3.76	1.00	1237456509851	9903
3.77	1.00	1237456509763	9904
3.80	1.00	1237456511664	9905
3.80	1.00	1237456510092	9906
3.88	1.00	1237456511082	9907
3.94	1.00	1237458232521	9908
4.05	1.00	1237456510624	9909
4.38	1.00	1237456509848	9910

PAR LES VARIABLES CONTINUES ACTIVES

COORD.	POIDS	LIBELLE DE LA VARIABLE	MOYENNE	ECART-TYPE	NUMERO
-0.65	9910.00	V035-descPhrasePosition.position:seuleInParagraphe	0.06	0.25	1
-0.58	9910.00	V110-zone.rubriqueName:	0.06	0.23	2
-0.52	9910.00	V136-title->niveau:0	0.07	0.25	3
-0.51	9910.00	V076-position.typeIndice:entiteNom;typePos:AMORCE	0.03	0.19	4
-0.46	9910.00	V108-encyclopedie.type:atlas	0.72	0.45	5
-0.45	9910.00	V022-position.typeIndice:entiteNom;typePos:IC	0.06	0.23	6
-0.28	9910.00	V055-dernierParag.position:finZone	0.06	0.23	7
-0.25	9910.00	V117-title->niveau:1	0.25	0.43	8
-0.24	9910.00	V091-premierParag.position:debutZone	0.01	0.12	9
-0.23	9910.00	V026-entiteNom.classe:plus	0.04	0.20	10
-0.22	9910.00	V094-zone.rubriqueName:Geo	0.18	0.39	11
-0.22	9910.00	V086-entiteNom.classe:mesure;sousClasse:fixe	0.06	0.31	12
-0.19	9910.00	V137-title->entiteNom.classe:lieu;sousClasse:pays	0.04	0.19	13
-0.18	9910.00	V100-premierParag.position:debutZoneSeul	0.03	0.18	14
-0.16	9910.00	V056-entiteNom.classe:mesure;sousClasse:indetermine	0.10	0.80	15
-0.14	9910.00	V078-entiteNom.classe:mesure;sousClasse:geopolitique	0.02	0.15	16
-0.13	9910.00	V047-position.typeIndice:entiteNom;typePos:END	0.13	0.34	17
-0.13	9910.00	V077-entiteNom.classe:mesure;sousClasse:evolutif	0.06	0.31	18
-0.11	9910.00	V142-title->position.typeIndice:entiteNom;typePos:AMORCE	0.00	0.06	19
-0.11	9910.00	V088-exprTemp.nature:ponctuel;sitTps:posteriorite	0.00	0.05	20
		Z O N E C E N T R A L E			
0.09	9910.00	V029-ptVue.type:prevision	0.02	0.16	127

0.09	9910.00	V009-argum.relation:temporelle	0.04	0.21	128
0.10	9910.00	V032-position.typeIndice:argum;typePos:IC	0.02	0.16	129
0.10	9910.00	V121-title->tpsVbx.temps:passeComp	0.00	0.06	130
0.10	9910.00	V017-argum.relation:consecution	0.04	0.20	131
0.10	9910.00	V120-title->entiteNom.classe:geopolitique	0.06	0.25	132
0.11	9910.00	V038-ptVue.type:recence	0.04	0.21	133
0.11	9910.00	V014-argum.relation:identite	0.03	0.17	134
0.12	9910.00	V053-premierParag.position:debutDivision	0.03	0.16	135
0.12	9910.00	V138-title->exprTemp.nature:ponctuel;sitTps:trespasse	0.03	0.16	136
0.12	9910.00	V016-descPhrasePosition.position:finParagraphe	0.16	0.36	137
0.14	9910.00	V015-argum.relation:correction	0.06	0.23	138
0.18	9910.00	V103-encyclopedie.type:GLI	0.09	0.29	139
0.21	9910.00	V042-dernierParag.position:finDivision	0.17	0.38	140
0.26	9910.00	V006-tpsVbx.temps:present	1.12	0.98	141
0.26	9910.00	V007-zone.rubriqueName:Eco	0.07	0.25	142
0.31	9910.00	V068-zone.rubriqueName:Hist	0.15	0.36	143
0.38	9910.00	V002-descPhraseType.type:assertion	0.97	0.17	144
0.39	9910.00	V003-encyclopedie.type:GUL	0.19	0.39	145
0.43	9910.00	V119-title->niveau:2	0.54	0.50	146

DESCRIPTION DU FACTEUR 2
PAR LES INDIVIDUS ACTIFS

COORD.	POIDS	IDENTIFICATEUR DE L'INDIVIDU	NUMERO
-3.19	1.00	1238262898562	1
-3.10	1.00	1238262902789	2
-3.05	1.00	1238262902790	3
-2.83	1.00	1238262902787	4
-2.82	1.00	1238262899157	5
-2.81	1.00	1238262900106	6
-2.80	1.00	1238262902715	7
-2.79	1.00	1238262899602	8
-2.77	1.00	1238262897190	9
-2.73	1.00	1238262902791	10
-2.73	1.00	1238262900107	11
-2.73	1.00	1238262900333	12
-2.72	1.00	1238262898641	13
-2.72	1.00	1238262899842	14
-2.71	1.00	1238262897702	15
-2.71	1.00	1238262902793	16
-2.71	1.00	1238262902788	17
-2.70	1.00	1238262902725	18
-2.67	1.00	1238262902731	19
-2.66	1.00	1238262898871	20
		Z O N E C E N T R A L E	
5.43	1.00	1237456510894	9891
5.51	1.00	1237456510631	9892
5.56	1.00	1237456510716	9893
5.57	1.00	1237458232636	9894
5.60	1.00	1237458232369	9895
5.66	1.00	1237456511637	9896
5.66	1.00	1237456510883	9897
5.66	1.00	1237456510639	9898
5.71	1.00	1237456510882	9899
5.76	1.00	1237458231894	9900
5.77	1.00	1237456510886	9901
5.96	1.00	1237456510881	9902
6.20	1.00	1237458232373	9903
6.25	1.00	1237458232180	9904
6.26	1.00	1237456510878	9905
6.46	1.00	1237456509949	9906
6.68	1.00	1237456510714	9907
6.75	1.00	1237458232381	9908
6.83	1.00	1237456511712	9909
8.94	1.00	1238262901418	9910

PAR LES VARIABLES CONTINUES ACTIVES

COORD.	POIDS	LIBELLE DE LA VARIABLE	MOYENNE	ECART-TYPE	NUMERO
-0.74	9910.00	V108-encyclopedie.type:atlas	0.72	0.45	1
-0.45	9910.00	V101-zone.rubriqueName:Med	0.17	0.38	2
-0.22	9910.00	V002-descPhraseType.type:assertion	0.97	0.17	3
-0.21	9910.00	V119-title->niveau:2	0.54	0.50	4
-0.16	9910.00	V006-tpsVbx.temps:present	1.12	0.98	5
-0.14	9910.00	V116-zone.rubriqueName:ArtLitt	0.03	0.18	6
-0.14	9910.00	V111-zone.rubriqueName:Sport	0.05	0.22	7
-0.14	9910.00	V099-zone.rubriqueName:ScTechn	0.13	0.33	8
-0.13	9910.00	V100-premierParag.position:debutZoneSeul	0.03	0.18	9

-0.12	9910.00	V042-dernierParag.position:finDivision	0.17	0.38	10
-0.09	9910.00	V128-title->entiteNom.classe:personne	0.05	0.24	11
-0.09	9910.00	V085-zone.rubriqueName:Soc	0.12	0.33	12
-0.08	9910.00	V092-zone.rubriqueName:FauneFlore	0.03	0.16	13
-0.07	9910.00	V066-premierParag.position:debutDivisionSeul	0.06	0.24	14
-0.07	9910.00	V135-title->entiteNom.classe:sigle	0.02	0.14	15
-0.07	9910.00	V138-title->exprTemp.nature:ponctuel;sitTps:trespasse	0.03	0.16	16
-0.05	9910.00	V039-argum.relation:justification	0.01	0.11	17
-0.05	9910.00	V018-argum.relation:explication	0.06	0.24	18
-0.05	9910.00	V033-argum.relation:illustration	0.01	0.10	19
-0.05	9910.00	V143-title->tpsVbx.temps:passeSimple	0.01	0.07	20
		Z O N E C E N T R A L E			
0.16	9910.00	V012-exprTemp.nature:ponctuel;sitTps:anteriorite	0.09	0.37	127
0.16	9910.00	V028-exprTemp.nature:ponctuel;sitTps:coincidence	0.04	0.23	128
0.17	9910.00	V077-entiteNom.classe:mesure;sousClasse:evolutif	0.06	0.31	129
0.17	9910.00	V021-entiteNom.classe:lieu;sousClasse:indetermine	0.14	0.51	130
0.17	9910.00	V062-entiteNom.classe:lieu;sousClasse:ville	0.10	0.56	131
0.21	9910.00	V076-position.typeIndice:entiteNom;typePos:AMORCE	0.03	0.19	132
0.22	9910.00	V022-position.typeIndice:entiteNom;typePos:IC	0.06	0.23	133
0.22	9910.00	V137-title->entiteNom.classe:lieu;sousClasse:pays	0.04	0.19	134
0.23	9910.00	V045-entiteNom.classe:lieu;sousClasse:pays	0.17	0.62	135
0.24	9910.00	V035-descPhrasePosition.position:seuleInParagraphe	0.06	0.25	136
0.25	9910.00	V094-zone.rubriqueName:Geo	0.18	0.39	137
0.26	9910.00	V047-position.typeIndice:entiteNom;typePos:END	0.13	0.34	138
0.27	9910.00	V053-premierParag.position:debutDivision	0.03	0.16	139
0.28	9910.00	V068-zone.rubriqueName:Hist	0.15	0.36	140
0.28	9910.00	V120-title->entiteNom.classe:geopolitique	0.06	0.25	141
0.33	9910.00	V010-entiteNom.classe:geopolitique	0.15	0.44	142
0.37	9910.00	V103-encyclopedie.type:GLI	0.09	0.29	143
0.37	9910.00	Obs	0.20	0.57	144
0.40	9910.00	V007-zone.rubriqueName:Eco	0.07	0.25	145
0.58	9910.00	V003-encyclopedie.type:GUL	0.19	0.39	146

DESCRIPTION DU FACTEUR 3
PAR LES INDIVIDUS ACTIFS

COORD.	POIDS	IDENTIFICATEUR DE L'INDIVIDU	NUMERO
-11.87	1.00	1238262902967	1
-9.15	1.00	1238262896802	2
-8.60	1.00	1238262896810	3
-8.50	1.00	1238262896806	4
-7.99	1.00	1238262902954	5
-7.35	1.00	1238262902932	6
-6.99	1.00	1238262899044	7
-6.41	1.00	1238262903380	8
-6.29	1.00	1238262903027	9
-5.88	1.00	1238262903006	10
-5.75	1.00	1237456510722	11
-5.57	1.00	1238262902887	12
-5.52	1.00	1238262902633	13
-5.43	1.00	1237456511644	14
-5.40	1.00	1237456510714	15
-5.39	1.00	1237456510725	16
-5.36	1.00	1237456510779	17
-5.29	1.00	1237456510084	18
-5.19	1.00	1237458232721	19
-5.14	1.00	1237458231959	20
		Z O N E C E N T R A L E	
6.04	1.00	1237458232703	9891
6.07	1.00	1237458232178	9892
6.09	1.00	1237456511712	9893
6.13	1.00	1238262901435	9894
6.15	1.00	1237458232195	9895
6.19	1.00	1238262900709	9896
6.20	1.00	1238262900450	9897
6.24	1.00	1237458232175	9898
6.26	1.00	1237458232189	9899
6.44	1.00	1238262896210	9900
6.77	1.00	1238262901368	9901
6.83	1.00	1238262901416	9902
6.86	1.00	1238262901074	9903
6.91	1.00	1238262900441	9904
7.15	1.00	1237458232191	9905
7.37	1.00	1238262900734	9906
7.38	1.00	1238262901418	9907
7.51	1.00	1237458232381	9908
7.93	1.00	1238262900739	9909
8.52	1.00	1238262900553	9910

PAR LES VARIABLES CONTINUES ACTIVES

COORD.	POIDS	LIBELLE DE LA VARIABLE	MOYENNE	ECART-TYPE	NUMERO
-0.36	9910.00	V068-zone.rubriqueName:Hist	0.15	0.36	1
-0.35	9910.00	V019-entiteNom.classe:personne	0.18	0.67	2
-0.33	9910.00	V023-exprTemp.nature:ponctuel;sitTps:trespasse	0.12	0.52	3
-0.32	9910.00	V012-exprTemp.nature:ponctuel;sitTps:anteriorite	0.09	0.37	4
-0.27	9910.00	V117-title->niveau:1	0.25	0.43	5
-0.22	9910.00	V052-position.typeIndice:exprTemp;typePos:IC	0.06	0.24	6
-0.21	9910.00	V044-position.typeIndice:exprTemp;typePos:END	0.05	0.22	7
-0.21	9910.00	V128-title->entiteNom.classe:personne	0.05	0.24	8
-0.20	9910.00	V138-title->exprTemp.nature:ponctuel;sitTps:trespasse	0.03	0.16	9
-0.17	9910.00	V116-zone.rubriqueName:ArtLitt	0.03	0.18	10
-0.17	9910.00	V110-zone.rubriqueName:	0.06	0.23	11
-0.16	9910.00	V136-title->niveau:0	0.07	0.25	12
-0.15	9910.00	V066-premierParag.position:debutDivisionSeul	0.06	0.24	13
-0.13	9910.00	V124-title->exprTemp.nature:ponctuel;sitTps:anteriorite	0.02	0.14	14
-0.12	9910.00	V053-premierParag.position:debutDivision	0.03	0.16	15
-0.12	9910.00	V134-title->ptVue.type:prevision	0.01	0.09	16
-0.12	9910.00	V049-exprTemp.nature:inachevee;sitTps:trespasse	0.01	0.11	17
-0.12	9910.00	V040-tpsVbx.temps:passeSimple	0.04	0.24	18
-0.11	9910.00	V106-position.typeIndice:exprTemp;typePos:AMORCE	0.00	0.06	19
-0.11	9910.00	V100-premierParag.position:debutZoneSeul	0.03	0.18	20
		Z O N E C E N T R A L E			
0.06	9910.00	V058-argum.relation:opposition	0.03	0.16	127
0.06	9910.00	V096-entiteNom.classe:mesure;sousClasse:estimation	0.00	0.05	128
0.07	9910.00	V092-zone.rubriqueName:FauneFlore	0.03	0.16	129
0.08	9910.00	V083-entiteNom.classe:lieu;sousClasse:riviere	0.02	0.19	130
0.09	9910.00	V085-zone.rubriqueName:Soc	0.12	0.33	131
0.09	9910.00	V034-exprTemp.nature:deictique;sitTps:coincidence	0.07	0.27	132
0.09	9910.00	V121-title->tpsVbx.temps:passeComp	0.00	0.06	133
0.10	9910.00	V108-encyclopedie.type:atlas	0.72	0.45	134
0.10	9910.00	V076-position.typeIndice:entiteNom;typePos:AMORCE	0.03	0.19	135
0.10	9910.00	V095-entiteNom.classe:lieu;sousClasse:pointCardinal	0.02	0.18	136
0.10	9910.00	V139-title->entiteNom.classe:lieu;sousClasse:ville	0.02	0.15	137
0.12	9910.00	V086-entiteNom.classe:mesure;sousClasse:fixe	0.06	0.31	138
0.23	9910.00	V078-entiteNom.classe:mesure;sousClasse:geopolitique	0.02	0.15	139
0.27	9910.00	V077-entiteNom.classe:mesure;sousClasse:evolutif	0.06	0.31	140
0.27	9910.00	V042-dernierParag.position:finDivision	0.17	0.38	141
0.37	9910.00	Obs	0.20	0.57	142
0.39	9910.00	V010-entiteNom.classe:geopolitique	0.15	0.44	143
0.39	9910.00	V119-title->niveau:2	0.54	0.50	144
0.49	9910.00	V120-title->entiteNom.classe:geopolitique	0.06	0.25	145
0.52	9910.00	V094-zone.rubriqueName:Geo	0.18	0.39	146

DESCRIPTION DU FACTEUR 4
PAR LES INDIVIDUS ACTIFS

COORD.	POIDS	IDENTIFICATEUR DE L'INDIVIDU	NUMERO
-11.30	1.00	1238262896802	1
-10.39	1.00	1238262896806	2
-10.28	1.00	1238262896783	3
-10.02	1.00	1238262896810	4
-8.42	1.00	1238262902967	5
-7.84	1.00	1238262901634	6
-7.38	1.00	1238262899044	7
-6.84	1.00	1238262902954	8
-6.58	1.00	1238262902764	9
-6.32	1.00	1238262902932	10
-6.20	1.00	1238262902765	11
-6.20	1.00	1238262897033	12
-6.17	1.00	1238262903671	13
-6.10	1.00	1238262902729	14
-5.99	1.00	1238262896958	15
-5.96	1.00	1237456511042	16
-5.95	1.00	1238262902763	17
-5.89	1.00	1238262897043	18
-5.87	1.00	1238262902769	19
-5.86	1.00	1238262902768	20
		Z O N E C E N T R A L E	
4.55	1.00	1238262897629	9891
4.57	1.00	1238262897630	9892
4.60	1.00	1238262897780	9893
4.62	1.00	1237456509926	9894
4.70	1.00	1237456509924	9895

4.72	1.00	1237456509918	9896
4.79	1.00	1237456509944	9897
4.82	1.00	1237456509934	9898
4.86	1.00	1237456509925	9899
4.90	1.00	1237456509928	9900
4.93	1.00	1237456509919	9901
4.97	1.00	1238262897626	9902
5.04	1.00	1238262897628	9903
5.08	1.00	1237456509933	9904
5.12	1.00	1238262899663	9905
5.16	1.00	1237456509920	9906
5.25	1.00	1237456509923	9907
5.37	1.00	1237456509931	9908
5.54	1.00	1237456509938	9909
5.77	1.00	1237456509937	9910

PAR LES VARIABLES CONTINUES ACTIVES

COORD.	POIDS	LIBELLE DE LA VARIABLE	MOYENNE	ECART-TYPE	NUMERO
-0.46	9910.00	V119-title->niveau:2	0.54	0.50	1
-0.39	9910.00	V138-title->exprTemp.nature:ponctuel;sitTps:trespasse	0.03	0.16	2
-0.34	9910.00	V128-title->entiteNom.classe:personne	0.05	0.24	3
-0.28	9910.00	V047-position.typeIndice:entiteNom;typePos:END	0.13	0.34	4
-0.27	9910.00	V023-exprTemp.nature:ponctuel;sitTps:trespasse	0.12	0.52	5
-0.27	9910.00	V116-zone.rubriqueName:ArtLitt	0.03	0.18	6
-0.26	9910.00	V022-position.typeIndice:entiteNom;typePos:IC	0.06	0.23	7
-0.25	9910.00	V111-zone.rubriqueName:Sport	0.05	0.22	8
-0.24	9910.00	V124-title->exprTemp.nature:ponctuel;sitTps:anteriorite	0.02	0.14	9
-0.24	9910.00	V019-entiteNom.classe:personne	0.18	0.67	10
-0.22	9910.00	V042-dernierParag.position:finDivision	0.17	0.38	11
-0.20	9910.00	V062-entiteNom.classe:lieu;sousClasse:ville	0.10	0.56	12
-0.20	9910.00	V068-zone.rubriqueName:Hist	0.15	0.36	13
-0.19	9910.00	V045-entiteNom.classe:lieu;sousClasse:pays	0.17	0.62	14
-0.17	9910.00	V108-encyclopedie.type:atlas	0.72	0.45	15
-0.17	9910.00	V056-entiteNom.classe:mesure;sousClasse:indetermine	0.10	0.80	16
-0.17	9910.00	V021-entiteNom.classe:lieu;sousClasse:indetermine	0.14	0.51	17
-0.15	9910.00	V076-position.typeIndice:entiteNom;typePos:AMORCE	0.03	0.19	18
-0.14	9910.00	V083-entiteNom.classe:lieu;sousClasse:riviere	0.02	0.19	19
-0.14	9910.00	V137-title->entiteNom.classe:lieu;sousClasse:pays	0.04	0.19	20
		Z O N E C E N T R A L E			
0.10	9910.00	V121-title->tpsVbx.temps:passeComp	0.00	0.06	127
0.10	9910.00	V038-ptVue.type:recence	0.04	0.21	128
0.10	9910.00	V048-position.typeIndice:ptVue;typePos:IC	0.01	0.09	129
0.11	9910.00	V034-exprTemp.nature:deictique;sitTps:coincidence	0.07	0.27	130
0.11	9910.00	V018-argum.relation:explication	0.06	0.24	131
0.11	9910.00	Obs	0.20	0.57	132
0.11	9910.00	V091-premierParag.position:debutZone	0.01	0.12	133
0.11	9910.00	V110-zone.rubriqueName:	0.06	0.23	134
0.12	9910.00	V029-ptVue.type:prevision	0.02	0.16	135
0.13	9910.00	V006-tpsVbx.temps:present	1.12	0.98	136
0.14	9910.00	V126-title->ptVue.type:distance	0.00	0.05	137
0.14	9910.00	V099-zone.rubriqueName:ScTechn	0.13	0.33	138
0.15	9910.00	V101-zone.rubriqueName:Med	0.17	0.38	139
0.16	9910.00	V055-dernierParag.position:finZone	0.06	0.23	140
0.19	9910.00	V136-title->niveau:0	0.07	0.25	141
0.20	9910.00	V100-premierParag.position:debutZoneSeul	0.03	0.18	142
0.24	9910.00	V066-premierParag.position:debutDivisionSeul	0.06	0.24	143
0.26	9910.00	V003-encyclopedie.type:GUL	0.19	0.39	144
0.30	9910.00	V007-zone.rubriqueName:Eco	0.07	0.25	145
0.46	9910.00	V117-title->niveau:1	0.25	0.43	146

DESCRIPTION DU FACTEUR 5
PAR LES INDIVIDUS ACTIFS

COORD.	POIDS	IDENTIFICATEUR DE L'INDIVIDU	NUMERO
-5.18	1.00	1238262901615	1
-4.77	1.00	1238262901623	2
-4.50	1.00	1238262901418	3
-4.48	1.00	1238262900610	4
-4.30	1.00	1238262901694	5
-4.23	1.00	1238262900532	6
-4.23	1.00	1238262900490	7
-4.17	1.00	1238262901618	8
-4.16	1.00	1238262901628	9
-4.15	1.00	1238262900611	10
-4.15	1.00	1238262901631	11
-4.13	1.00	1238262901625	12
-4.09	1.00	1238262901138	13

-4.05	1.00	1238262901624	14
-4.05	1.00	1238262901135	15
-4.04	1.00	1238262901134	16
-4.03	1.00	1238262900607	17
-4.03	1.00	1238262901136	18
-3.98	1.00	1238262901614	19
-3.98	1.00	1238262900606	20
		Z O N E C E N T R A L E	
6.79	1.00	1238262897499	9891
6.81	1.00	1238262897188	9892
6.99	1.00	1238262895538	9893
7.01	1.00	1238262898437	9894
7.04	1.00	1238262897501	9895
7.06	1.00	1238262901326	9896
7.39	1.00	1238262896198	9897
7.45	1.00	1238262902560	9898
7.92	1.00	1237458232636	9899
8.41	1.00	1238262897781	9900
8.41	1.00	1238262897783	9901
8.43	1.00	1238262897782	9902
9.15	1.00	1238262897785	9903
9.69	1.00	1238262897780	9904
9.76	1.00	1238262897784	9905
9.83	1.00	1238262897500	9906
11.46	1.00	1238262902111	9907
11.59	1.00	1238262902109	9908
11.80	1.00	1238262902110	9909
12.08	1.00	1238262902112	9910

PAR LES VARIABLES CONTINUES ACTIVES

COORD.	POIDS	LIBELLE DE LA VARIABLE	MOYENNE	ECART-TYPE	NUMERO
-0.60	9910.00	V117-title->niveau:1	0.25	0.43	1
-0.42	9910.00	V066-premierParag.position:debutDivisionSeul	0.06	0.24	2
-0.24	9910.00	V137-title->entiteNom.classe:lieu;sousClasse:pays	0.04	0.19	3
-0.23	9910.00	V094-zone.rubriqueName:Geo	0.18	0.39	4
-0.11	9910.00	V108-encyclopedie.type:atlas	0.72	0.45	5
-0.10	9910.00	V142-title->position.typeIndice:entiteNom;typePos:AMORCE	0.00	0.06	6
-0.09	9910.00	V139-title->entiteNom.classe:lieu;sousClasse:ville	0.02	0.15	7
-0.08	9910.00	V062-entiteNom.classe:lieu;sousClasse:ville	0.10	0.56	8
-0.08	9910.00	V141-title->entiteNom.classe:lieu;sousClasse:riviere	0.00	0.07	9
-0.06	9910.00	V053-premierParag.position:debutDivision	0.03	0.16	10
-0.06	9910.00	V055-dernierParag.position:finZone	0.06	0.23	11
-0.05	9910.00	V149-title->exprTemp.nature:inachevee;sitTps:anteriorite	0.00	0.04	12
-0.05	9910.00	V072-tpsVbx.temps:passeAnt	0.00	0.07	13
-0.05	9910.00	V130-title->ptVue.type:recence	0.01	0.11	14
-0.05	9910.00	V086-entiteNom.classe:mesure;sousClasse:fixe	0.06	0.31	15
-0.05	9910.00	V088-exprTemp.nature:ponctuel;sitTps:posteriorite	0.00	0.05	16
-0.05	9910.00	V076-position.typeIndice:entiteNom;typePos:AMORCE	0.03	0.19	17
-0.04	9910.00	V135-title->entiteNom.classe:sigle	0.02	0.14	18
-0.04	9910.00	V129-title->niveau:3	0.12	0.32	19
-0.04	9910.00	V045-entiteNom.classe:lieu;sousClasse:pays	0.17	0.62	20
		Z O N E C E N T R A L E			
0.10	9910.00	V038-ptVue.type:recence	0.04	0.21	127
0.10	9910.00	V081-ptVue.type:source	0.00	0.05	128
0.10	9910.00	V005-tpsVbx.temps:passeComp	0.27	0.51	129
0.11	9910.00	V015-argum.relation:correction	0.06	0.23	130
0.11	9910.00	V032-position.typeIndice:argum;typePos:IC	0.02	0.16	131
0.11	9910.00	V048-position.typeIndice:ptVue;typePos:IC	0.01	0.09	132
0.13	9910.00	V026-entiteNom.classe:plus	0.04	0.20	133
0.13	9910.00	Obs	0.20	0.57	134
0.14	9910.00	V145-title->tpsVbx.temps:futur	0.00	0.05	135
0.14	9910.00	V134-title->ptVue.type:prevision	0.01	0.09	136
0.15	9910.00	V091-premierParag.position:debutZone	0.01	0.12	137
0.15	9910.00	V140-title->tpsVbx.temps:imparfait	0.00	0.03	138
0.16	9910.00	V103-encyclopedie.type:GLI	0.09	0.29	139
0.16	9910.00	V123-title->exprTemp.nature:deictique;sitTps:coincidence	0.00	0.05	140
0.16	9910.00	V147-title->argum.relation:correction	0.00	0.02	141
0.19	9910.00	V110-zone.rubriqueName:	0.06	0.23	142
0.22	9910.00	V119-title->niveau:2	0.54	0.50	143
0.24	9910.00	V143-title->tpsVbx.temps:passeSimple	0.01	0.07	144
0.58	9910.00	V100-premierParag.position:debutZoneSeul	0.03	0.18	145
0.66	9910.00	V136-title->niveau:0	0.07	0.25	146

DESCRIPTION DU FACTEUR 6
PAR LES INDIVIDUS ACTIFS

COORD.	POIDS	IDENTIFICATEUR DE L'INDIVIDU	NUMERO
-7.52	1.00	1237456509811	1
-7.52	1.00	1237456509807	2
-7.05	1.00	1238262896783	3
-6.77	1.00	1238262896837	4
-6.77	1.00	1237456509803	5
-6.51	1.00	1237456509850	6
-6.50	1.00	1237456509854	7
-6.46	1.00	1237456509853	8
-6.25	1.00	1237456509855	9
-6.18	1.00	1237456509852	10
-6.08	1.00	1237456509856	11
-5.95	1.00	1237456509857	12
-5.91	1.00	1237456509932	13
-5.81	1.00	1237456509858	14
-5.63	1.00	1237456509766	15
-5.60	1.00	1237456509859	16
-5.55	1.00	1237456509925	17
-5.48	1.00	1237456509926	18
-5.43	1.00	1237456509924	19
-5.39	1.00	1237456509922	20
		Z O N E C E N T R A L E	
5.48	1.00	1238262903027	9891
5.61	1.00	1238262895762	9892
5.62	1.00	1238262897413	9893
5.62	1.00	1238262896290	9894
5.65	1.00	1238262903678	9895
5.67	1.00	1238262896341	9896
5.70	1.00	1238262901062	9897
5.80	1.00	1238262902186	9898
5.86	1.00	1238262903458	9899
6.00	1.00	1238262901623	9900
6.12	1.00	1238262902110	9901
6.16	1.00	1238262902967	9902
6.25	1.00	1238262897436	9903
6.29	1.00	1238262902112	9904
6.36	1.00	1238262903385	9905
6.38	1.00	1238262901032	9906
6.95	1.00	1237456510714	9907
7.33	1.00	1238262899044	9908
7.60	1.00	1237458232636	9909
7.77	1.00	1238262903380	9910

PAR LES VARIABLES CONTINUES ACTIVES

COORD.	POIDS	LIBELLE DE LA VARIABLE	MOYENNE	ECART-TYPE	NUMERO
-0.41	9910.00	V007-zone.rubriqueName:Eco	0.07	0.25	1
-0.29	9910.00	V124-title->exprTemp.nature:ponctuel;sitTps:anteriorite	0.02	0.14	2
-0.28	9910.00	V003-encyclopedie.type:GUL	0.19	0.39	3
-0.21	9910.00	V121-title->tpsVbx.temps:passeComp	0.00	0.06	4
-0.20	9910.00	V128-title->entiteNom.classe:personne	0.05	0.24	5
-0.19	9910.00	V129-title->niveau:3	0.12	0.32	6
-0.19	9910.00	V118-title->tpsVbx.temps:present	0.03	0.18	7
-0.19	9910.00	V111-zone.rubriqueName:Sport	0.05	0.22	8
-0.17	9910.00	V022-position.typeIndice:entiteNom;typePos:IC	0.06	0.23	9
-0.17	9910.00	V126-title->ptVue.type:distance	0.00	0.05	10
-0.14	9910.00	V035-descPhrasePosition.position:seuleInParagraphe	0.06	0.25	11
-0.13	9910.00	V125-title->exprTemp.nature:ponctuel;sitTps:coincidence	0.01	0.09	12
-0.12	9910.00	V036-descPhraseType.type:interrogation	0.00	0.05	13
-0.11	9910.00	V122-title->argum.relation:temporelle	0.00	0.04	14
-0.10	9910.00	V076-position.typeIndice:entiteNom;typePos:AMORCE	0.03	0.19	15
-0.09	9910.00	V110-zone.rubriqueName:	0.06	0.23	16
-0.08	9910.00	V101-zone.rubriqueName:Med	0.17	0.38	17
-0.08	9910.00	V138-title->exprTemp.nature:ponctuel;sitTps:trespasse	0.03	0.16	18
-0.08	9910.00	V056-entiteNom.classe:mesure;sousClasse:indetermine	0.10	0.80	19
-0.08	9910.00	V123-title->exprTemp.nature:deictique;sitTps:coincidence	0.00	0.05	20
		Z O N E C E N T R A L E			
0.12	9910.00	V099-zone.rubriqueName:ScTechn	0.13	0.33	127
0.14	9910.00	V068-zone.rubriqueName:Hist	0.15	0.36	128
0.15	9910.00	V083-entiteNom.classe:lieu;sousClasse:riviere	0.02	0.19	129
0.15	9910.00	V043-exprTemp.nature:inachevee;sitTps:anteriorite	0.02	0.14	130
0.16	9910.00	V100-premierParag.position:debutZoneSeul	0.03	0.18	131
0.17	9910.00	V108-encyclopedie.type:atlas	0.72	0.45	132
0.17	9910.00	V117-title->niveau:1	0.25	0.43	133
0.18	9910.00	V094-zone.rubriqueName:Geo	0.18	0.39	134
0.19	9910.00	V002-descPhraseType.type:assertion	0.97	0.17	135
0.20	9910.00	V077-entiteNom.classe:mesure;sousClasse:evolutif	0.06	0.31	136

0.21	9910.00	V005-tpsVbx.temps:passeComp	0.27	0.51	137
0.22	9910.00	V021-entiteNom.classe:lieu;sousClasse:indetermine	0.14	0.51	138
0.22	9910.00	V012-exprTemp.nature:ponctuel;sitTps:anteriorite	0.09	0.37	139
0.24	9910.00	V034-exprTemp.nature:deictique;sitTps:coincidence	0.07	0.27	140
0.25	9910.00	V028-exprTemp.nature:ponctuel;sitTps:coincidence	0.04	0.23	141
0.26	9910.00	Obs	0.20	0.57	142
0.26	9910.00	V066-premierParag.position:debutDivisionSeul	0.06	0.24	143
0.28	9910.00	V045-entiteNom.classe:lieu;sousClasse:pays	0.17	0.62	144
0.29	9910.00	V044-position.typeIndice:exprTemp;typePos:END	0.05	0.22	145
0.32	9910.00	V052-position.typeIndice:exprTemp;typePos:IC	0.06	0.24	146

DESCRIPTION DU FACTEUR 7
PAR LES INDIVIDUS ACTIFS

COORD.	POIDS	IDENTIFICATEUR DE L'INDIVIDU	NUMERO
-12.13	1.00	1237458231959	1
-8.65	1.00	1237458231956	2
-8.43	1.00	1237458231950	3
-8.33	1.00	1237458231938	4
-8.27	1.00	1237458231951	5
-8.20	1.00	1237458232636	6
-7.93	1.00	1238262896783	7
-7.34	1.00	1237458231944	8
-7.23	1.00	1237458231943	9
-7.23	1.00	1237458231952	10
-7.23	1.00	1237458231941	11
-7.23	1.00	1237458231953	12
-6.93	1.00	1238262902478	13
-6.68	1.00	1237458231939	14
-6.42	1.00	1237458231958	15
-6.37	1.00	1238262897043	16
-6.26	1.00	1237458231937	17
-6.26	1.00	1238262897039	18
-6.23	1.00	1237458231940	19
-6.22	1.00	1238262897033	20
		Z O N E C E N T R A L E	
6.12	1.00	1238262900034	9891
6.24	1.00	1238262900425	9892
6.42	1.00	1238262901615	9893
6.49	1.00	1238262901323	9894
6.55	1.00	1238262901327	9895
6.56	1.00	1238262901324	9896
6.56	1.00	1238262901329	9897
6.60	1.00	1238262901328	9898
6.86	1.00	1238262901325	9899
6.90	1.00	1237456511222	9900
7.00	1.00	1238262901321	9901
7.33	1.00	1238262901201	9902
7.61	1.00	1238262902564	9903
8.54	1.00	1238262900640	9904
11.66	1.00	1238262901322	9905
13.36	1.00	1238262902110	9906
13.77	1.00	1238262902112	9907
13.94	1.00	1238262902111	9908
14.47	1.00	1238262902109	9909
14.66	1.00	1238262901032	9910

PAR LES VARIABLES CONTINUES ACTIVES

COORD.	POIDS	LIBELLE DE LA VARIABLE	MOYENNE	ECART-TYPE	NUMERO
-0.34	9910.00	V119-title->niveau:2	0.54	0.50	1
-0.27	9910.00	V106-position.typeIndice:exprTemp;typePos:AMORCE	0.00	0.06	2
-0.26	9910.00	V020-entiteNom.classe:sigle	0.09	0.38	3
-0.23	9910.00	V042-dernierParag.position:finDivision	0.17	0.38	4
-0.23	9910.00	V035-descPhrasePosition.position:seuleInParagraphe	0.06	0.25	5
-0.22	9910.00	V099-zone.rubriqueName:ScTechn	0.13	0.33	6
-0.21	9910.00	V028-exprTemp.nature:ponctuel;sitTps:coincidence	0.04	0.23	7
-0.18	9910.00	V081-ptVue.type:source	0.00	0.05	8
-0.17	9910.00	V052-position.typeIndice:exprTemp;typePos:IC	0.06	0.24	9
-0.16	9910.00	V110-zone.rubriqueName:	0.06	0.23	10
-0.15	9910.00	V085-zone.rubriqueName:Soc	0.12	0.33	11
-0.14	9910.00	V048-position.typeIndice:ptVue;typePos:IC	0.01	0.09	12
-0.13	9910.00	V056-entiteNom.classe:mesure;sousClasse:indetermine	0.10	0.80	13
-0.12	9910.00	V055-dernierParag.position:finZone	0.06	0.23	14
-0.11	9910.00	V027-tpsVbx.temps:conditionnel	0.02	0.15	15
-0.11	9910.00	V054-exprTemp.nature:ponctuel;sitTps:indetermine	0.01	0.08	16
-0.11	9910.00	V041-position.typeIndice:ptVue;typePos:END	0.01	0.09	17
-0.11	9910.00	V038-ptVue.type:recence	0.04	0.21	18

-0.11	9910.00	V077-entiteNom.classe:mesure;sousClasse:evolutif	0.06	0.31	19
-0.10	9910.00	V135-title->entiteNom.classe:sigle	0.02	0.14	20
		Z O N E C E N T R A L E			
0.08	9910.00	V016-descPhrasePosition.position:finParagraphe	0.16	0.36	127
0.08	9910.00	V001-descPhrasePosition.position:debutParagraphe	0.16	0.37	128
0.09	9910.00	V086-entiteNom.classe:mesure;sousClasse:fixe	0.06	0.31	129
0.09	9910.00	V068-zone.rubriqueName:Hist	0.15	0.36	130
0.10	9910.00	V047-position.typeIndice:entiteNom;typePos:END	0.13	0.34	131
0.11	9910.00	V131-title->entiteNom.classe:lieu;sousClasse:indetermine	0.02	0.13	132
0.14	9910.00	V145-title->tpsVbx.temps:futur	0.00	0.05	133
0.15	9910.00	V006-tpsVbx.temps:present	1.12	0.98	134
0.17	9910.00	V062-entiteNom.classe:lieu;sousClasse:ville	0.10	0.56	135
0.19	9910.00	V136-title->niveau:0	0.07	0.25	136
0.19	9910.00	V045-entiteNom.classe:lieu;sousClasse:pays	0.17	0.62	137
0.20	9910.00	V083-entiteNom.classe:lieu;sousClasse:riviere	0.02	0.19	138
0.21	9910.00	V147-title->argum.relation:correction	0.00	0.02	139
0.24	9910.00	V139-title->entiteNom.classe:lieu;sousClasse:ville	0.02	0.15	140
0.24	9910.00	V002-descPhraseType.type:assertion	0.97	0.17	141
0.27	9910.00	V095-entiteNom.classe:lieu;sousClasse:pointCardinal	0.02	0.18	142
0.28	9910.00	V129-title->niveau:3	0.12	0.32	143
0.32	9910.00	V143-title->tpsVbx.temps:passeSimple	0.01	0.07	144
0.34	9910.00	V094-zone.rubriqueName:Geo	0.18	0.39	145
0.41	9910.00	V100-premierParag.position:debutZoneSeul	0.03	0.18	146

DESCRIPTION DU FACTEUR 8
PAR LES INDIVIDUS ACTIFS

COORD.	POIDS	IDENTIFICATEUR DE L'INDIVIDU	NUMERO
-12.32	1.00	1237458231950	1
-9.98	1.00	1237458231959	2
-7.82	1.00	1237458231956	3
-7.77	1.00	1237458231818	4
-7.75	1.00	1237458231951	5
-7.48	1.00	1237458231938	6
-7.41	1.00	1237458231936	7
-7.19	1.00	1237458231958	8
-6.98	1.00	1237458231943	9
-6.98	1.00	1237458231952	10
-6.98	1.00	1237458231941	11
-6.89	1.00	1237458231937	12
-6.75	1.00	1237458231942	13
-6.75	1.00	1237458231945	14
-6.73	1.00	1237458231954	15
-6.73	1.00	1237458231957	16
-6.63	1.00	1237458231946	17
-6.54	1.00	1237458231953	18
-6.48	1.00	1237458231940	19
-6.45	1.00	1237458231960	20
		Z O N E C E N T R A L E	
5.60	1.00	1238262902662	9891
5.60	1.00	1238262902791	9892
5.67	1.00	1238262902478	9893
5.69	1.00	1238262902785	9894
5.69	1.00	1238262902787	9895
5.77	1.00	1238262902729	9896
5.90	1.00	1238262902792	9897
6.03	1.00	1238262902715	9898
6.12	1.00	1238262902760	9899
6.25	1.00	1238262902720	9900
6.27	1.00	1238262902794	9901
6.31	1.00	1238262902790	9902
6.35	1.00	1238262900450	9903
6.39	1.00	1238262902724	9904
6.54	1.00	1237456510304	9905
6.98	1.00	1238262902573	9906
7.26	1.00	1238262902663	9907
7.53	1.00	1238262902712	9908
7.60	1.00	1238262902967	9909
7.90	1.00	1238262902789	9910

PAR LES VARIABLES CONTINUES ACTIVES

COORD.	POIDS	LIBELLE DE LA VARIABLE	MOYENNE	ECART-TYPE	NUMERO
-0.47	9910.00	V103-encyclopedie.type:GLI	0.09	0.29	1
-0.34	9910.00	V129-title->niveau:3	0.12	0.32	2
-0.28	9910.00	V106-position.typeIndice:exprTemp;typePos:AMORCE	0.00	0.06	3

-0.18	9910.00	V099-zone.rubriqueName:ScTechn	0.13	0.33	4
-0.18	9910.00	V052-position.typeIndice:exprTemp;typePos:IC	0.06	0.24	5
-0.14	9910.00	V053-premierParag.position:debutDivision	0.03	0.16	6
-0.13	9910.00	V001-descPhrasePosition.position:debutParagraphe	0.16	0.37	7
-0.11	9910.00	V105-exprTemp.nature:duree;sitTps:trespasse	0.00	0.01	8
-0.10	9910.00	V068-zone.rubriqueName:Hist	0.15	0.36	9
-0.08	9910.00	V054-exprTemp.nature:ponctuel;sitTps:indetermine	0.01	0.08	10
-0.07	9910.00	V101-zone.rubriqueName:Med	0.17	0.38	11
-0.06	9910.00	V063-exprTemp.nature:anaphorique;sitTps:indetermine	0.01	0.13	12
-0.06	9910.00	V020-entiteNom.classe:sigle	0.09	0.38	13
-0.06	9910.00	V111-zone.rubriqueName:Sport	0.05	0.22	14
-0.06	9910.00	V064-exprTemp.nature:duree;sitTps:indetermine	0.01	0.11	15
-0.05	9910.00	V005-tpsVbx.temps:passeComp	0.27	0.51	16
-0.05	9910.00	V085-zone.rubriqueName:Soc	0.12	0.33	17
-0.04	9910.00	V067-zone.rubriqueName:Droit	0.00	0.04	18
-0.04	9910.00	V139-title->entiteNom.classe:lieu;sousClasse:ville	0.02	0.15	19
-0.04	9910.00	V104-position.typeIndice:argum;typePos:END	0.00	0.03	20
		Z O N E C E N T R A L E			
0.14	9910.00	V021-entiteNom.classe:lieu;sousClasse:indetermine	0.14	0.51	127
0.15	9910.00	V108-encyclopedie.type:atlas	0.72	0.45	128
0.15	9910.00	V084-ptVue.type:definition	0.00	0.05	129
0.15	9910.00	V126-title->ptVue.type:distance	0.00	0.05	130
0.15	9910.00	V010-entiteNom.classe:geopolitique	0.15	0.44	131
0.16	9910.00	V029-ptVue.type:prevision	0.02	0.16	132
0.17	9910.00	V032-position.typeIndice:argum;typePos:IC	0.02	0.16	133
0.17	9910.00	V013-tpsVbx.temps:futur	0.03	0.20	134
0.17	9910.00	V134-title->ptVue.type:prevision	0.01	0.09	135
0.18	9910.00	V007-zone.rubriqueName:Eco	0.07	0.25	136
0.18	9910.00	V015-argum.relation:correction	0.06	0.23	137
0.18	9910.00	V003-encyclopedie.type:GUL	0.19	0.39	138
0.21	9910.00	V128-title->entiteNom.classe:personne	0.05	0.24	139
0.21	9910.00	V117-title->niveau:1	0.25	0.43	140
0.21	9910.00	V038-ptVue.type:recence	0.04	0.21	141
0.22	9910.00	V048-position.typeIndice:ptVue;typePos:IC	0.01	0.09	142
0.23	9910.00	V118-title->tpsVbx.temps:present	0.03	0.18	143
0.24	9910.00	V019-entiteNom.classe:personne	0.18	0.67	144
0.32	9910.00	V138-title->exprTemp.nature:ponctuel;sitTps:trespasse	0.03	0.16	145
0.39	9910.00	V116-zone.rubriqueName:ArtLitt	0.03	0.18	146

C.5 Règles créées sur la base des axes 2 à 7 de l'ACP

```
-- facteur 2~: 6 non ob sur 32
SELECT distinct d1.Id, d1.Obs, d1.Text
FROM dataAnalysis d1, dataAnalysis d2, dataAnalysis d3, dataAnalysis d7
where (d1.`V007-zone.rubriqueName:Eco`~!= 0
or d1.`V068-zone.rubriqueName:Hist`~!= 0
or d1.`V094-zone.rubriqueName:Geo`~!= 0)
and d2.`V010-entiteNom.classe:geopolitique`~!= 0
and d3.`V120-title->entiteNom.classe:geopolitique`~!= 0
and d7.`V045-entiteNom.classe:lieu;sousClasse:pays`~!= 0
and d1.Id = d2.Id
and d2.Id = d3.Id
and d3.Id = d7.Id
UNION
-- facteur 3~: 3 non ob sur 65
SELECT distinct d1.Id, d1.Obs, d1.Text
FROM dataAnalysis d1, dataAnalysis d2, dataAnalysis d3, dataAnalysis d4, dataAnalysis d5
where d1.`V094-zone.rubriqueName:Geo`~!= 0
and d2.`V120-title->entiteNom.classe:geopolitique`~!= 0
and d3.`V010-entiteNom.classe:geopolitique`~!= 0
and d4.`V119-title->niveau:2`~!= 0
and (d5.`V077-entiteNom.classe:mesure;sousClasse:evolutif`~!= 0 or d5.`V078-entiteNom.classe:mesure;sousClasse:geopolitique`~!= 0)
and d1.Id = d2.Id
and d2.Id = d3.Id
and d3.Id = d4.Id
and d4.Id = d5.Id
UNION
-- facteur 4~: 15 non ob sur 31
SELECT distinct d1.Id, d1.Obs, d1.Text
FROM dataAnalysis d1,  dataAnalysis d5, dataAnalysis d6
where (d1.`V007-zone.rubriqueName:Eco`~!= 0
or d1.`V110-zone.rubriqueName:`~!= 0
or d1.`V099-zone.rubriqueName:ScTechn`~!= 0
or d1.`V101-zone.rubriqueName:Med`~!= 0 )
and (d5.`V029-ptVue.type:prevision`~!= 0
or d5.`V038-ptVue.type:recence`~!= 0)
and d6.`V034-exprTemp.nature:deictique;sitTps:coincidence`~!= 0
and d1.Id = d5.Id
and d5.Id = d6.Id
UNION
-- facteur 5~: 13 sur 22
SELECT distinct d1.Id, d1.Obs, d1.Text
FROM dataAnalysis d1, dataAnalysis d2,  dataAnalysis d4
where (d2.`V100-premierParag.position:debutZoneSeul`~!= 0
or d2.`V091-premierParag.position:debutZone`~!= 0)
and (d4.`V147-title->argum.relation:correction`~!= 0
or d4.`V123-title->exprTemp.nature:deictique;sitTps:coincidence`~!= 0
or d4.`V134-title->ptVue.type:prevision`~!= 0)
and d1.Id = d2.Id
and d2.Id = d4.Id
UNION
-- facteur 6~: 15 sur 86
SELECT distinct d1.Id, d1.Obs, d1.Text
FROM dataAnalysis d1, dataAnalysis d2, dataAnalysis d4
where d2.`V077-entiteNom.classe:mesure;sousClasse:evolutif`~!= 0
and (d4.`V028-exprTemp.nature:ponctuel;sitTps:coincidence`~!= 0
or d4.`V034-exprTemp.nature:deictique;sitTps:coincidence`~!= 0)
and d1.Id = d2.Id
and d2.Id = d4.Id
UNION
-- facteur 6~: 6 sur 29
SELECT distinct d1.Id, d1.Obs, d1.Text
FROM dataAnalysis d1, dataAnalysis d2, dataAnalysis d4
where (d1.`V045-entiteNom.classe:lieu;sousClasse:pays`~!= 0
or d1.`V021-entiteNom.classe:lieu;sousClasse:indetermine`~!= 0)
and d2.`V077-entiteNom.classe:mesure;sousClasse:evolutif`~!= 0
and (d4.`V028-exprTemp.nature:ponctuel;sitTps:coincidence`~!= 0
or d4.`V034-exprTemp.nature:deictique;sitTps:coincidence`~!= 0)
and d1.Id = d2.Id
and d2.Id = d4.Id
UNION
-- facteur 7~: 8 sur 28
SELECT distinct d1.Id, d1.Obs, d1.Text
FROM dataAnalysis d1, dataAnalysis d2, dataAnalysis d4, dataAnalysis d5
where d1.`V020-entiteNom.classe:sigle`~!= 0
and (d2.`V056-entiteNom.classe:mesure;sousClasse:indetermine`~!= 0
or d2.`V077-entiteNom.classe:mesure;sousClasse:evolutif`~!= 0 )
and (d4.`V028-exprTemp.nature:ponctuel;sitTps:coincidence`~!= 0
or d4.`V034-exprTemp.nature:deictique;sitTps:coincidence`~!= 0 )
and d1.Id = d2.Id
and d2.Id = d4.Id
```

```
and d4.Id = d5.Id
UNION
-- facteur 7~: 14 sur 28
SELECT distinct d1.Id, d1.Obs, d1.Text
FROM dataAnalysis d1, dataAnalysis d4, dataAnalysis d5, dataAnalysis d3
where (d3.`V006-tpsVbx.temps:present`~!= 0
or d3.`V013-tpsVbx.temps:futur`~!= 0
or d3.`V027-tpsVbx.temps:conditionnel`~!= 0)
and (d4.`V028-exprTemp.nature:ponctuel;sitTps:coincidence`~!= 0
or d4.`V034-exprTemp.nature:deictique;sitTps:coincidence`~!= 0
or d4.`V093-exprTemp.nature:deictique;sitTps:posteriorite`~!= 0
or d4.`V088-exprTemp.nature:ponctuel;sitTps:posteriorite`~!= 0 )
and d1.Id = d4.Id
and d4.Id = d5.Id
and d5.Id = d3.Id
```

Annexe D

Résultats règles Fouille de texte

D.1 corpusApprComplet

```
root@dellriou:/home/marion/Documents/THESE/statistiques/Binarisation# /usr/local/bin/mvclassif.sh
/usr/local/bin/mvminer2bdfree.sh 1 corpusComplet 2 1 5 0.001
<experience>
  <params proto="/usr/local/bin/mvminer2bdfree.sh" ref="corpusComplet" nclass="2" u="1" delta="5" minsup="0.001" />
/usr/local/bin/mvminer2bdfree.sh corpusComplet.train1 1 1 5 0.001 > corpusComplet.train1.o
awk -f /usr/local/bin/mvmkrules.awk corpusComplet.train1 2 0.001 8924
awk -f /usr/local/bin/mvnegclass.awk corpusComplet.train1 2 >  corpusComplet.train1.neg
/usr/local/bin/mvminer2bdfree.sh corpusComplet.train1.neg -1 -1 5 0.001 > corpusComplet.train1.neg.o
mvclassifrc.sh 1 corpusComplet 2
--> mvclassrc.sh corpusComplet.train1 2
mvscorule corpusComplet.train1.o.cs1.cover corpusComplet.train1 | sed 's/^://'| /mvcolmed
mediane 26.7631
mvscorule corpusComplet.train1.o.cs2.cover corpusComplet.train1 | sed 's/^://'| /mvcolmed
mediane 21.9433
#class preci rap fscore confusion
1 89.5 94.9 92.1 798 43
2 57 37.7 45.4 94 57

1.5 73.2 66.3 68.8 446 50 ~: moyenne
----------------------------------
#class preci rap fscore confusion
1 94.9 70.7 81.1 595 246
2 32.6 78.8 46.1 32 119

1.5 63.7 74.8 63.6 314 182 ~: moyenne
----------------------------------
#class preci rap fscore confusion
1 94.9 71.1 81.3 598 243
2 32.9 78.8 46.4 32 119

1.5 63.9 75 63.8 315 181 ~: moyenne
----------------------------------
----------------------------------
  <score ref="corpusComplet" rc="corpusComplet" cs="86" cn="71" cscn="72" />
  <stats positives="10630-1571-" pos-cover="4604-1378-" negatives="5990-15578-" neg-cover="3041-5335-" />
  <time>138</time>
</experience>

aire sous la courbe ROC~: 79.2261
```

D.2 corpusApprIPSeuls

```
root@dellriou:/home/marion/Documents/THESE/statistiques/Binarisation# /usr/local/bin/mvclassif.sh
/usr/local/bin/mvminer2bdfree.sh 1 corpusApprIPseuls 2 1 5 0.001
<experience>
  <params proto="/usr/local/bin/mvminer2bdfree.sh" ref="corpusApprIPseuls" nclass="2" u="1" delta="5" minsup="0.001" />
/usr/local/bin/mvminer2bdfree.sh corpusApprIPseuls.train1 1 1 5 0.001 > corpusApprIPseuls.train1.o
awk -f /usr/local/bin/mvmkrules.awk corpusApprIPseuls.train1 2 0.001 7697
awk -f /usr/local/bin/mvnegclass.awk corpusApprIPseuls.train1 2 >  corpusApprIPseuls.train1.neg
/usr/local/bin/mvminer2bdfree.sh corpusApprIPseuls.train1.neg -1 -1 5 0.001 > corpusApprIPseuls.train1.neg.o
mvclassifrc.sh 1 corpusApprIPseuls 2
--> mvclassrc.sh corpusApprIPseuls.train1 2
mvscorule corpusApprIPseuls.train1.o.cs1.cover corpusApprIPseuls.train1 | sed 's/^://'| /mvcolmed
mediane 3.49265
mvscorule corpusApprIPseuls.train1.o.cs2.cover corpusApprIPseuls.train1 | sed 's/^://'| /mvcolmed
mediane 9.48599
#class preci rap fscore confusion
1 85 96.2 90.2 684 27
2 48.1 17.1 25.3 121 25

1.5 66.5 56.7 57.7 402 26 ~: moyenne
----------------------------------
#class preci rap fscore confusion
1 87.3 86.9 87.1 618 93
2 37.6 38.4 38 90 56

1.5 62.4 62.6 62.5 354 74.5 ~: moyenne
----------------------------------
#class preci rap fscore confusion
1 87.1 87.6 87.4 623 88
2 38 37 37.5 92 54

1.5 62.6 62.3 62.4 358 71 ~: moyenne
----------------------------------
----------------------------------
  <score ref="corpusApprIPseuls" rc="corpusApprIPseuls" cs="82" cn="78" cscn="78" />
  <stats positives="1067-190-" pos-cover="962-190-" negatives="291-1280-" neg-cover="289-1111-" />
  <time>15</time>
</experience>

aire sous la courbe ROC~: 61.5042
```

D.3 corpusApprIPHierar

```
root@dellriou:/home/marion/Documents/THESE/statistiques/Binarisation# /usr/local/bin/mvclassif.sh
/usr/local/bin/mvminer2bdfree.sh 1 corpusApprIPHierar 2 1 5 0.001
<experience>
  <params proto="/usr/local/bin/mvminer2bdfree.sh" ref="corpusApprIPHierar" nclass="2" u="1" delta="5" minsup="0.001" />
/usr/local/bin/mvminer2bdfree.sh corpusApprIPHierar.train1 1 1 5 0.001 > corpusApprIPHierar.train1.o
awk -f /usr/local/bin/mvmkrules.awk corpusApprIPHierar.train1 2 0.001 7751
awk -f /usr/local/bin/mvnegclass.awk corpusApprIPHierar.train1 2 >  corpusApprIPHierar.train1.neg
/usr/local/bin/mvminer2bdfree.sh corpusApprIPHierar.train1.neg -1 -1 5 0.001 > corpusApprIPHierar.train1.neg.o
mvclassifrc.sh 1 corpusApprIPHierar 2
--> mvclassrc.sh corpusApprIPHierar.train1 2
mvscorule corpusApprIPHierar.train1.o.cs1.cover corpusApprIPHierar.train1 | sed 's/^://'| /mvcolmed
mediane 3.82828
mvscorule corpusApprIPHierar.train1.o.cs2.cover corpusApprIPHierar.train1 | sed 's/^://'| /mvcolmed
mediane 11.8553
#class preci rap fscore confusion
1 86.6 95.8 91 685 30
2 57.7 27.9 37.6 106 41

1.5 72.2 61.8 64.3 396 35.5 ~: moyenne
---------------------------------
#class preci rap fscore confusion
1 88.4 85.5 86.9 611 104
2 39.2 45.6 42.1 80 67

1.5 63.8 65.5 64.5 346 85.5 ~: moyenne
---------------------------------
#class preci rap fscore confusion
1 88.5 85.9 87.2 614 101
2 39.9 45.6 42.5 80 67

1.5 64.2 65.7 64.8 347 84 ~: moyenne
---------------------------------
---------------------------------
  <score ref="corpusApprIPHierar" rc="corpusApprIPHierar" cs="84" cn="78" cscn="79" />
  <stats positives="1414-363-" pos-cover="1191-358-" negatives="469-1830-" neg-cover="462-1454-" />
  <time>22</time>
</experience>

aire sous la courbe ROC~: 66.1591
```

D.4 corpusApprIPPos

```
root@dellriou:/home/marion/Documents/THESE/statistiques/Binarisation# /usr/local/bin/mvclassif.sh
/usr/local/bin/mvminer2bdfree.sh 1 corpusApprIPPos 2 1 5 0.001
<experience>
  <params proto="/usr/local/bin/mvminer2bdfree.sh" ref="corpusApprIPPos" nclass="2" u="1" delta="5" minsup="0.001" />
/usr/local/bin/mvminer2bdfree.sh corpusApprIPPos.train1 1 1 5 0.001 > corpusApprIPPos.train1.o
awk -f /usr/local/bin/mvmkrules.awk corpusApprIPPos.train1 2 0.001 8517
awk -f /usr/local/bin/mvnegclass.awk corpusApprIPPos.train1 2 >  corpusApprIPPos.train1.neg
/usr/local/bin/mvminer2bdfree.sh corpusApprIPPos.train1.neg -1 -1 5 0.001 > corpusApprIPPos.train1.neg.o
mvclassifrc.sh 1 corpusApprIPPos 2
--> mvclassrc.sh corpusApprIPPos.train1 2
mvscorule corpusApprIPPos.train1.o.cs1.cover corpusApprIPPos.train1 | sed 's/^://'| /mvcolmed
mediane 8.25628
mvscorule corpusApprIPPos.train1.o.cs2.cover corpusApprIPPos.train1 | sed 's/^://'| /mvcolmed
mediane 17.9696
#class preci rap fscore confusion
1 87.1 96.7 91.7 771 26
2 58.1 24 34 114 36

1.5 72.6 60.4 62.8 442 31 ~: moyenne
---------------------------------
#class preci rap fscore confusion
1 90.7 78 83.9 622 175
2 33 57.3 41.8 64 86

1.5 61.8 67.7 62.9 343 130 ~: moyenne
---------------------------------
#class preci rap fscore confusion
1 90.6 78.5 84.1 626 171
2 33.2 56.7 41.9 65 85

1.5 61.9 67.6 63 346 128 ~: moyenne
---------------------------------
---------------------------------
  <score ref="corpusApprIPPos" rc="corpusApprIPPos" cs="85" cn="74" cscn="75" />
  <stats positives="3730-599-" pos-cover="2634-579-" negatives="1800-5206-" neg-cover="1187-3131-" />
  <time>56</time>
</experience>

aire sous la courbe ROC~: 68.63
```

D.5 corpusApprEpure

```
root@dellriou:/home/marion/Documents/THESE/statistiques/Binarisation# /usr/local/bin/mvclassif.sh
/usr/local/bin/mvminer2bdfree.sh 1 corpusApprEpure 2 1 5 0.001
<experience>
  <params proto="/usr/local/bin/mvminer2bdfree.sh" ref="corpusApprEpure" nclass="2" u="1" delta="5" minsup="0.001" />
/usr/local/bin/mvminer2bdfree.sh corpusApprEpure.train1 1 1 5 0.001 > corpusApprEpure.train1.o
awk -f /usr/local/bin/mvmkrules.awk corpusApprEpure.train1 2 0.001 8918
awk -f /usr/local/bin/mvnegclass.awk corpusApprEpure.train1 2 >  corpusApprEpure.train1.neg
/usr/local/bin/mvminer2bdfree.sh corpusApprEpure.train1.neg -1 -1 5 0.001 > corpusApprEpure.train1.neg.o
mvclassifrc.sh 1 corpusApprEpure 2
--> mvclassrc.sh corpusApprEpure.train1 2
mvscorule corpusApprEpure.train1.o.cs1.cover corpusApprEpure.train1 | sed 's/^://'| /mvcolmed
mediane 13.8554
mvscorule corpusApprEpure.train1.o.cs2.cover corpusApprEpure.train1 | sed 's/^://'| /mvcolmed
mediane 22.2892
#class preci rap fscore confusion
1 88.6 95.8 92.1 806 35
2 57.3 31.1 40.3 104 47

1.5 72.9 63.5 66.2 455 41 ~: moyenne
---------------------------------
#class preci rap fscore confusion
1 92.5 82 87 690 151
2 38.6 62.9 47.9 56 95

1.5 65.6 72.5 67.4 373 123 ~: moyenne
---------------------------------
#class preci rap fscore confusion
1 92.4 82.3 87 692 149
2 38.7 62.3 47.7 57 94

1.5 65.5 72.3 67.4 374 122 ~: moyenne
---------------------------------
---------------------------------
  <score ref="corpusApprEpure" rc="corpusApprEpure" cs="85" cn="79" cscn="79" />
  <stats positives="3224-638-" pos-cover="2207-602-" negatives="1110-4043-" neg-cover="993-2582-" />
  <time>46</time>
</experience>

aire sous la courbe ROC~: 72.96
```

Annexe E

Format des données de test

	corpus Test Atlas	corpus Test GLI	corpus Test Entier
mots	82 388	17096	99484
phrases	3 377	539	3916
adverbiaux temporels	1581	329	1910
entités nommées dans un titre	337	33	370
entités nommées dans un paragraphe	3420	704	4124
périphrases verbales	31	3	34
expressions de point de vue	203	71	274
temps verbaux	4826	944	5770
type de phrase	3220	529	3749
position des phrases dans le paragraphe	1205	267	1472
position des indices dans la phrase	1307	196	1503
premier paragraphe	110	39	149
dernier paragraphe	107	31	138
Total	16347	3146	19493

FIG. E.1 - *Tous les indices génériques par corpus*