Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs

Marion Laignelet* — **François Rioult****

* *Laboratoire CLLE-ERSS - UMR 5263 CNRS*
allées Antonio Machado, 31058 Toulouse Cedex
marion.laignelet@univ-tlse2.fr

** *Université de Caen Basse-Normandie, Laboratoire GREYC - UMR 6072 CNRS*
Campus 2 Côte de Nacre, 14032 Caen cedex
Francois.Rioult@unicaen.fr

RÉSUMÉ. *Cet article vise la description et le repérage automatique de segments contenant de l'obsolescence dans les documents de type encyclopédique. Nous supposons, malgré le caractère non linguistique de ce phénomène, que des indices discursifs permettent le repérage de ces segments. Nous travaillons sur un corpus annoté manuellement par des experts sur lequel nous projetons des indices repérés automatiquement. Nous utilisons des techniques d'apprentissage automatique pour évaluer le pouvoir prédictif de nos indices. À l'aide de techniques de classification supervisée, nous montrons que nos hypothèses sont pertinentes et permettent d'envisager le déploiement d'une méthode automatique pour l'aide au repérage de segments obsolescents.*

ABSTRACT. *This paper deals with the description and the automatic tracking of text segments containing obsolescence in encyclopedia texts. We assume that despite the non-linguistic nature of this phenomenon, discursive cues are relevant to track those segments. For that purpose, we have worked on a corpus which has been manually annotated by experts and on which we have projected automatically tracked cues. We use machine learning techniques to evaluate the predictive power of our cues. We show, using supervised classification, that our hypotheses enable us to build an automatic procedure to assist human experts.*

MOTS-CLÉS : *repérage automatique de l'obsolescence, indices discursifs, textes encyclopédiques, classification supervisée, aire sous la courbe ROC.*

KEYWORDS: *automatic tracking of obsolescence, discursive cues, encyclopedic texts, supervised classification, area under the ROC curve.*

1. Introduction

Cet article s'inscrit dans le cadre d'un projet de recherche visant à la fois la description mais également le repérage automatique de segments obsolescents dans des documents de type encyclopédique. La visée applicative de ce travail consiste en la création d'un outil d'aide à la mise à jour de textes encyclopédiques dans le domaine de l'édition. Nous proposons de traiter la problématique de la mise à jour de l'information à travers la notion d'obsolescence : un segment obsolescent est défini comme un segment textuel contenant de l'information susceptible d'évolution dans le temps.

Nous travaillons sur la base d'un corpus composé de textes encyclopédiques issus du monde éditorial : ces textes publiés doivent ou devront faire l'objet de mises à jour concrètes à court et long terme. L'obsolescence est alors définie selon ce besoin de mise à jour.

Malgré le caractère intrinsèquement non linguistique de ce phénomène, nous supposons que des indices sémantiques et discursifs peuvent permettre le repérage des segments d'obsolescence. Le système mis en œuvre et décrit ici a pour objectif le repérage des indices et combinaisons d'indices à même de devenir des marqueurs de l'obsolescence. Ce travail s'appuie sur les recherches menées autour des notions de marqueurs textuels et discursifs comme les mots-repères ou les mots-titres, notions déjà envisagées par Edmondson (1997), les « cue phrases » (Grosz et Sidner, 1986) ou encore les éléments participant de l'analyse de la structure de texte (Marcu, 2000). Considérant le caractère multifonctionnel des marqueurs de surface (Grosz et Sidner, 1986), nous nous focalisons sur leur fonction pragmatique, c'est-à-dire leur aptitude à déterminer un segment d'information évolutive. Les aspects discursifs des documents à travers les titres (Ho-Dac *et al.*, 2004), les cadres de discours (Charolles, 1997) ou encore la position dans le document occupent une place importante dans nos recherches.

Pour évaluer la pertinence de notre approche, nous utilisons une technique d'apprentissage automatique, la classification supervisée (Sebastiani, 2002). Cette démarche permet à la fois de vérifier si oui ou non une machine peut repérer les segments obsolescents à l'aide de nos indices et de valider leur intérêt. L'apprentissage fournit également un modèle des données, par exemple sous forme de règles, que l'expert peut analyser, dans une optique de découverte de connaissance, puis discuter, dans le but d'intégrer sa propre connaissance métier.

La première section est consacrée à la description de notre corpus annoté manuellement, des segments contenant de l'information obsolescente et de la définition de la notion d'obsolescence (section 2). Nous décrivons ensuite quels indices linguistiques (sémantiques et discursifs) sont utilisés (section 3). Dans la section 4, nous présentons la méthode d'apprentissage automatique mise en œuvre. La section 5 rend compte de la performance de notre méthode selon plusieurs paramètres : tout d'abord, une comparaison par rapport à des approches de base, puis une évaluation de l'apport des différents niveaux d'indices linguistiques considérés, enfin une évaluation de notre classifieur à travers le jugement humain : les règles apprises sont projetées sur un nou-

veau corpus et les segments d'obsolescence repérés automatiquement sont finalement évalués par les experts. Nous discutons enfin dans la section 7 des connaissances linguistiques, sous forme de règles (combinaisons d'indices linguistiques) renvoyées par le classifieur.

2. Données et phénomène observé

Cette section rassemble les éléments nécessaires à la compréhension de l'article. Nous commençons par décrire les données textuelles à notre disposition, puis nous expliquons le phénomène d'obsolescence.

2.1. *Le corpus d'apprentissage*

Comme nous l'avons mentionné dans l'introduction de cet article, le corpus sur lequel nous travaillons est composé de textes venant du monde professionnel, édités et publiés dans le domaine public. Les textes sélectionnés, issus de deux bases principales proviennent de deux sous-corpus : ATLAS, qui regroupe des fiches encyclopédiques éditées par les éditions Atlas, et LAROUSSE qui est constitué d'entrées extraites d'encyclopédies des éditions Larousse (le *Grand Larousse Informatisé* et le *Grand Universel Larousse*). Ce corpus d'apprentissage compte environ 10 000 phrases.

La principale caractéristique de ce corpus est de regrouper des textes de type encyclopédique : de fait, ils sont classés en fonction du domaine auquel ils se rapportent : géographie, histoire, faune et flore, sciences et techniques, sport, économie, etc.

2.2. *L'obsolescence, description générale*

L'obsolescence est un phénomène non linguistique, créé par l'usage : les segments d'obsolescence se définissent d'abord par rapport à un besoin réel, à savoir la mise à jour éditoriale. L'information qu'ils contiennent est susceptible d'évolution, de modification ou de changement dans le temps. Nous ne cherchons pas à repérer de manière automatique l'évolution de la connaissance à proprement parler. C'est à travers la prise en compte de signaux textuels présents dans les textes qu'un segment sera ou non considéré comme potentiellement obsolescent. Par ailleurs, ces signaux ne sont pas forcément posés intentionnellement par le rédacteur de l'article encyclopédique.

L'exemple 1 présente un segment d'obsolescence. L'auteur y exprime une issue possible et probable concernant les recherches sur le sida. On observe dans cet extrait la présence d'indices temporels mais également celle d'indices modaux qui s'avèrent également centraux dans notre interprétation de l'obsolescence.

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*

Le corpus d'apprentissage a été annoté manuellement selon le caractère obsolescent ou non des segments [1] qui le composent : un expert linguiste [2] a annoté le corpus ATLAS et le corpus LAROUSSE qui a été également annoté par trois experts rédacteurs [3]. Comme le montrent les chiffres du tableau 1, la part de l'obsolescence est de 15 % environ dans le corpus d'apprentissage.

	ATLAS	**LAROUSSE**	Corpus d'apprentissage
Nombre total de phrases	7 142	2 874	9 916
Nombre de phrases obsolescentes	927	581	1 508
Pourcentage de phrases obsolescentes	12,9 %	20,2 %	15,2 %

Tableau 1. *Proportion de segments obsolescents dans le corpus d'apprentissage*

La différence de proportion d'obsolescence entre le sous-corpus ATLAS et le sous-corpus LAROUSSE vient du fait que les textes du sous-corpus ATLAS ont été réunis selon les données rendues disponibles par l'éditeur : il contient proportionnellement plus de fiches appartenant au domaine de la géographie, domaine qui se trouve être également très enclin au besoin de mise à jour (en géographie, 27,7 % des phrases sont potentiellement à mettre à jour alors qu'en histoire, leur proportion n'est que de 8,1 %). Le sous-corpus LAROUSSE est quant à lui plus homogène.

Une première analyse des segments annotés manuellement nous amène à distinguer deux grandes classes d'obsolescence : d'un côté les segments dont l'information est devenue fausse (la connaissance est envisagée dans un moment T ; à $T+1$,

1. L'unité prise en compte est la phrase.
2. Moi-même, Marion Laignelet.
3. Rédacteurs des éditions Larousse.

elle est susceptible d'avoir évolué) ; de l'autre, ceux, dont l'information n'est plus pertinente (d'un point de vue informationnel) au moment où elle est lue. Comparons les deux exemples construits suivants :

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

(2) *En 2004, le PIB par habitant de la France est de 27 600 euros.*

Dans l'exemple (1), sachant qu'on est actuellement en 2010, et que le lecteur va naturellement interpréter l'adverbial *aujourd'hui* comme étant l'année en cours, l'information est fausse puisque le PIB de la France le plus actuel (chiffres de 2008) est de 33 800 dollars [4]. À l'inverse, l'exemple (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 vraissemblablement actualiser à la fois la référence temporelle (*En 2010*) et la valeur du PIB associée.

Dans les deux cas, il s'agit de segments d'obsolescence. Ces exemples montrent la diversité des informations susceptibles d'évolution dans les textes et de fait la diversité des marques linguistiques impliquées pour les repérer de manière automatique.

2.3. *L'obsolescence, un phénomène consensuel*

Parce que nous disposons d'une multi-annotation humaine [5], il nous a été possible d'évaluer le taux d'accord de jugement sur l'obsolescence. Nous avons mesuré les taux de recouvrement des annotations manuelles. Il en ressort que le sous-corpus LAROUSSE est en moyenne composé de 10 à 15 % de segments obsolètes par annotateur et que l'accord observé entre chacun de ces juges se situe entre 87 et 92 %.

Le coefficient Kappa est traditionnellement utilisé pour évaluer les degrés d'accord entre juges. Concernant l'obsolescence, le taux d'accord est situé entre 0,35 et 0,50 : ce score très faible est directement lié à la forte disproportion des classes (15 % de segments obsolescents contre 85 % de segments non obsolescents). L'accord entre nos juges est mieux traduit par le coefficient r de Finn [6] (Hripcsak et Heitjan, 2002). Ce coefficient permet d'aplanir la disproportion des classes en comparant la proportion des accords observés à une situation aléatoire considérant, dans notre cas bien précis, que chaque annotateur a une chance sur deux de déclarer un segment obsolescent (en situation de hasard). Les scores pour le coefficient 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).

4. La situation serait sans doute identique avec une phrase ne contenant pas d'adverbial temporel. En effet, lorsqu'il n'y a pas de référence temporelle précise, le temps verbal donne certaines indications : ici le présent suggère une interprétation déictique. Mais dans les corpus ce n'est pas toujours le cas (présent historique par exemple).

5. Le sous-corpus LAROUSSE a été annoté par quatre experts différents.

6. Nous utilisons l'algorithme existant dans le logiciel R.

Ces résultats montrent tout d'abord qu'il n'y a pas une grande variation de jugement entre les quatre 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 cela montre 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 qui fera mieux que l'humain.

3. Repérage automatique des indices sémantiques et discursifs

Cette section précise la démarche linguistique employée pour repérer de façon automatique des indices sémantiques et discursifs potentiellement pertinents pour la caractérisation de l'obsolescence.

3.1. *Méthodologie*

Notre intuition et notre compétence de linguiste nous ont naturellement amenés à orienter nos recherches sur les indices de type temporel comme marqueurs potentiels de l'obsolescence. De plus, notre intérêt de longue date pour les travaux en discours nous a entraînés vers l'exploitation d'indices à plus grande granularité tels que les titres ou la position des paragraphes au sein des documents. Enfin, à la lumière des annotations fournies par les experts, nous avons élargi les types d'indices vers la prise en compte des entités nommées, des valeurs chiffrées ou encore des expressions du point de vue du locuteur comme indicateurs de l'obsolescence.

Le corpus d'apprentissage fait émerger 150 indices différents. Leur repérage et leur annotation sémantique sont effectués de manière entièrement automatique avec l'outil ALIDIS [7] développé à l'aide de la plate-forme LinguaStream [8] (Widlocher et Bilhaut, 2005).

Cet outil nous a permis d'utiliser et de développer des modules de traitement de la langue dédiés à des problématiques spécifiques : segmentation en mots, étiquetage morphosyntaxique (appel de l'outil TreeTagger), projection de lexiques [9], exploitation du balisage XML, macro-expressions régulières pour le repérage de la position des mots et des groupes de mots dans les phrases, grammaires ProLog pour le repérage du temps, de la modalité, des expressions de point de vue, des périphrases verbales ou encore des entités nommées.

7. Pour : Annotation LInguistique de DIScours.

8. http ://www.linguastream.org

9. Les lexiques ont été constitués sur la base d'études linguistiques pour ce qui est des lexiques de temps, de préposition, etc. ou sont le résultat de transformation d'autres données (exploitation de l'encyclopédie Larousse pour les lexiques de noms propres ou de sigles, par exemple).

Nous avons fait le choix de créer nos propres modules de repérage des entités nommées (et du temps également) pour deux raisons principales. Tout d'abord, nos besoins en termes de repérage et d'annotation sémantique des indices sont assez spécifiques : par exemple, pour le traitement du temps (repérage et annotation sémantique), nous n'avons besoin que de deux types d'information, le type de découpage temporel et la nature de la référence temporelle qui est calculée automatiquement en fonction de l'expression reconnue. La seconde raison pour laquelle nous avons développé nos propres outils est purement technique : il est souvent délicat de faire communiquer des outils et/ou des bases de connaissances lorsqu'ils sont développés sous des environnements différents et sans forcément tenir compte de leur réutilisabilité ou de leur adaptabilité dans des systèmes plus larges ou des plates-formes de traitement. Le fonctionnement des modules de repérage que nous avons créés (constitutifs de l'outil ALIDIS) est expliqué dans Laignelet (2009) et les ressources et programmes [10] sont disponibles en ligne [11].

Voyons le détail des grandes classes d'indices linguistiques que nous prenons en considération.

3.2. *Description des indices potentiels de l'obsolescence (repérés automatiquement par l'outil* ALIDIS*)*

En plus de leur diversité, ces indices présentent la caractéristique d'être multi-échelle, c'est-à-dire d'apparaître à différents niveaux textuels.

Les indices de type syntagmatique sont les plus nombreux et sont sémantiquement très variés.

Une première classe d'indices concerne les informations temporelles. Le temps joue effectivement un rôle prépondérant dans les segments d'obsolescence. L'analyseur temporel mis en œuvre repère et annote sémantiquement la grande classe des adverbiaux temporels : syntagmes prépositionnels (*dans les années trente*, *de 1980 à 2000*), adverbes (*aujourd'hui*), syntagmes nominaux (*les années 20*). La sémantique temporelle mise en œuvre exploite un découpage temporel extrêmement simplifié mais suffisant pour la tâche visée. En effet, notre tâche ne nécessite pas, du moins dans un premier temps, un modèle temporel qui rende compte de manière exhaustive du découpage temporel du texte, de l'ancrage des événements dans une référence temporelle ou encore de la succession des événements sur la ligne du temps comme l'envisage Reichenbach (1966) ou Gosselin (2005). L'analyseur temporel renvoit deux types d'information différents.

Tout d'abord, la nature de l'expression qui est évaluée selon les valeurs suivantes : anaphorique si l'expression temporelle doit être calculée en fonction du contexte (*trois*

10. Il s'agit essentiellement de grammaires Prolog, de macro-expressions régulières et de lexiques au format XML.
11. `http://marion.laignelet.free.fr`

jours avant), déictique si la date doit être calculée selon le moment d'énonciation (*aujourd'hui*), de durée si l'expression exprime une durée (*en trente ans*), de type itération si le processus est itératif (*tous les ans*), ponctuel (*En 2008*) et enfin elle peut être de type inachevé lorsque la frontière finale de l'intervalle n'est potentiellement pas refermée (*depuis 1997*).

Le second trait concerne le découpage temporel. Nous ne cherchons ni à effectuer un calcul précis sur les événements d'un texte, ni à calculer leur enchaînement. Pour chacune des expressions temporelles repérées, nous calculons les cinq valeurs suivantes, identifiées relativement aux besoins en termes de pratique éditoriale : la valeur antériorité++ pour les dates antérieures à 1949, antériorité pour les dates de 1950 à 1989, coïncidence pour les dates de 1990 à 2008, postériorité pour les dates après 2009 et indéterminé pour les expressions qu'on ne peut pas calculer, comme les anaphoriques.

En relation avec les notions de temps, nous prenons également en compte les indices aspectuels et modaux à travers, entre autres, le repérage de périphrases verbales et celui des temps verbaux. Ainsi, nous exploitons les expressions verbales présentant une action dont l'accomplissement débute, est en cours ou achevé. Par exemple, l'expression *des recherches sont en cours* est annotée comme une périphrase verbale dont l'accomplissement est en cours.

Les entités nommées semblent également jouer un rôle important dans les segments d'obsolescence. Ce que nous entendons par entité nommée est relativement vaste : nous y englobons des expressions de mesure (*130 hab./km*^{2}), de lieu (*à Paris*), de personne (des noms propres principalement), des sigles, des noms d'organisation, etc.

Enfin, une dernière grande classe d'indices qui semble susceptible de marquer l'obsolescence concerne les expressions exprimant un point de vue. Le point de vue peut prendre différentes valeurs selon que l'énonciateur se distancie des propos qu'il tient, se les approprie, les juge importants, nouveaux, etc. Par exemple, le syntagme prépositionnel *Selon les estimations de l'INSEE* est annoté comme étant de type source puisqu'une source précise est donnée, l'expression *on suppose* est annotée comme étant de type jugement, le syntagme nominal *les nouveaux formats de compression des données* est annoté comme étant de type récence du fait de la présence de l'adjectif *nouveau*.

L'ensemble de ces indices de type syntagmatique peuvent être réalisés au sein de l'unité phrase : dans ce cas, nous parlons d'indices intraphrastiques. Ils peuvent également apparaître dans des titres : nous parlons alors d'indices hiérarchiques. Les indices hiérarchiques correspondent au fait qu'une phrase peut être sous la dépendance d'une autre unité, le titre. Dans notre protocole d'annotation, un titre ne peut pas faire partie d'un segment d'obsolescence. En revanche, il peut être un bon prédicteur d'obsolescence pour les phrases qui sont sous sa dépendance. Ainsi, les indices que nous repérons dans les phrases sont également repérés dans les titres ; chaque phrase hérite ensuite des indices présents dans le titre qui la gouverne.

Parallèlement aux indices intraphrastiques et hiérarchiques, nous traitons les indices positionnels. Ils peuvent être de deux types. Les indices positionnels phrastiques rendent compte de la position des indices intraphrastiques au sein de l'unité phrase (début et fin de phrase). Les indices positionnels textuels rendent compte de la position des unités phrase au sein des paragraphes (première phrase ou dernière phrase du paragraphe) et des unités paragraphe au sein du document (premier paragraphe ou dernier paragraphe de la section, sous section, etc.) ou encore du niveau de hiérarchie dans le document (par exemple les niveaux des titres).

Les indices externes concernent par exemple le type de document ou le domaine pour lequel le document est rédigé. Nous exploitons une dizaine de rubriques différentes (histoire, géographie, faune et flore, etc).

Nous considérons tous ces indices et les relations potentielles entre ces indices de manière aveugle et sans *a priori* sur leur fonctionnement. Ainsi, même si nous nous attendons à observer des associations d'indices propres à un genre (certains indices et/ou configurations d'indices pertinentes en histoire mais pas en géographie), nous préférons, dans un premier temps du moins, laisser le texte et les statistiques nous le dévoiler. Par ailleurs, l'idée que le texte puisse révéler l'obsolescence à travers sa propre progression argumentative ou sa progression aspecto-temporelle (par exemple considérer les changements de temps, ou les ruptures thématiques) est une piste de recherche qui nous semble prometteuse et vers laquelle nous souhaitons aller à terme mais pour laquelle nous n'avons, pour le moment, ni solution de modélisation cohérente ni solution d'implémentation satisfaisante.

La performance de ces modules de repérage automatique a été évaluée manuellement sur un dixième du corpus d'apprentissage. Les résultats sont donnés dans le tableau 2 : pour chaque classe d'indices sont indiqués la précision (proportion d'indices correctement retrouvés) et le rappel (proportion d'indices retrouvés).

	Précision	Rappel
Temps verbaux	97 %	98 %
Adverbiaux temporels	92 %	98 %
Périphrases verbales	99 %	43 %
Entités nommées	99 %	83 %
Expression du point de vue	73 %	98 %
Moyenne	93 %	85 %

Tableau 2. *Performance globale de l'outil* ALIDIS

On observe quelques disparités. Tout d'abord, le repérage des périphrases verbales (*des recherches sont en cours*, *les essais sont terminés*) a une précision correcte mais un rappel médiocre (43 %) : nous préférons repérer moins d'expressions de ce type mais être sûrs de la qualité de celles qui sont bien repérées. Dans une moindre mesure, la situation est identique pour le repérage des entités nommées. Concernant les expressions du point de vue, le rappel est correct mais la précision est moyenne (73 %) :

le repérage automatique de telles expressions est plus délicat notamment parce qu'il nécessite en lui-même une prise en compte plus large du contexte. Par exemple, nous souhaitons repérer une expression comme *la recherche [prévoit] des avancées considérables dans ce domaine* mais pas une expression comme *la loi [prévoit] un an de prison pour...* ; or le fait que nous nous basions sur la présence du verbe *prévoir* nous renvoie les deux cas. Des règles contextuelles pourraient sans doute permettre de lever ce type d'ambiguïté.

D'une manière générale, ces résultats sont suffisants pour envisager un traitement statistique à grande échelle, du moins dans un premier temps. Peut-être sera-t-il, à terme, nécessaire de pallier ces disparités en pondérant chacun des indices ou classes d'indices en fonction de leur fiabilité.

3.3. *Vers des configurations d'indices*

Notre objectif est de mettre en place des techniques d'analyse de données 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 (*cf.* la démarche classificatoire émergente de Biber (1989)).

Pour répondre à ce double objectif, nous avons mis en place, d'un côté une méthode de type descriptif (statistiques de base et analyse en composantes principales), et de l'autre une méthode de type prédictif (apprentissage automatique).

Les statistiques de base et l'analyse en composantes principales (ACP) 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 (c'est-à-dire des indices) selon leur pertinence et leur « significativité » au sein des segments d'obsolescence.

L'objectif du processus d'apprentissage automatique 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).

3.3.1. *Statistiques descriptives : pertinence des indices*

Afin de savoir si les indices linguistiques que nous prenons en compte sont significatifs dans les segments obsolescents, nous avons, dans un premier temps, effectué un calcul de la corrélation entre chacun des indices linguistiques et la nature obso-

lescente ou non du segment considéré [12]. Ces premières statistiques nous apprennent deux choses. D'abord, les indices considérés sont effectivement pertinents pour notre tâche, que ce soit parce qu'ils sont fortement corrélés à l'obsolescence ou parce qu'ils en sont indépendants. Ces tests confirment également l'idée que l'obsolescence ne peut être appréhendée par des marqueurs univoques ou référant à des indices isolés.

Les cinquante et un indices les plus corrélés positivement à la variable $[obsol]$ [13] ne permettent cependant pas de repérer efficacement les segments obsolescents : si l'on considère que ces indices sont des marqueurs de l'obsolescence et que l'on fait l'hypothèse que leur présence entraîne l'obsolescence de la phrase, alors un tel système aurait de mauvaises performances : 18 % de précision et 95 % de rappel. Utilisés seuls, les indices considérés ne sont pas suffisants. De plus, la nature même de certains indices, notamment ceux qui sont positionnels (phrastiques, de paragraphe ou de document), n'ont que peu de sens s'ils sont utilisés seuls. Nous supposons donc que c'est en termes de configurations que les indices deviendront de bons marqueurs de l'obsolescence. Pour tester ce point, nous avons mis en place une analyse de données : l'objectif principal est de faire émerger des corrélations complexes entre les indices.

3.3.2. *Analyse en composantes principales : corrélations entre indices*

L'analyse en composantes principales (Lebart *et al.*, 1995) met en lumière les tendances fortes de corrélations entre plusieurs indices linguistiques et la nature obsolescente du segment (dans ce qui est appelé des composantes principales).

Dans l'axe 3 de l'analyse, on observe une corrélation entre les indices temporels référant à une date proche du moment d'énonciation, certaines entités nommées (de mesure ou de type géopolitique) et la variable $[obsol]$. Cette combinaison est également corrélée aux variables référant aux domaines de l'économie, de la géographie et de l'histoire. L'exemple 2 illustre ce cas.

> 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 de l'ACP : 1238262900553

Exemple 2. *Les entités nommées de types mesure et géopolitique dans un segment d'obsolescence*

Dans l'axe 2, les indices de temps, les entités nommées, les indices de positions textuelles particulières (fin de paragraphe et/ou fin de section) et la variable $obsol$ sont corrélés.

12. Les variables sont quantitatives, elles sont évaluées en nombre d'occurences de l'indice/variable et le nombre d'annotations manuelles pour un même segment.
13. $[obsol]$ pour obsolescence.

Parce que aucune composante (axe) ne se détache clairement des autres, l'ACP montre que le phénomène de l'obsolescence est un problème non trivial et qu'il n'est pas possible de chercher à l'appréhender avec peu d'indices ou des configurations d'indices simples (Laignelet et Rioult, 2009).

Ces résultats confirment néanmoins la nécessité de mettre au jour des marqueurs complexes pour l'obsolescence (c'est-à-dire constitués d'indices sémantiques variés et multi-échelles). Nous mettons en œuvre un apprentissage automatique des combinaisons présentes dans les segments d'obsolescence. La section suivante présente le système d'apprentissage automatique mis en place.

4. Apprentissage automatique supervisé

Cette section met à l'épreuve la pertinence de nos indices linguistiques pour prédire automatiquement l'obsolescence d'un segment. Nous utilisons une méthode d'apprentissage pour :

– *(i)* évaluer numériquement la pertinence de nos indices ;

– *(ii)* obtenir des connaissances sur le pouvoir prédictif de certaines configurations d'indices ;

– *(iii)* proposer une solution d'automatisation performante de la tâche de reconnaissance de l'obsolescence à intégrer dans un processus éditorial.

Les méthodes d'apprentissage sont appliquées sur le corpus d'apprentissage présenté dans la section 2.1. Sur ce corpus d'apprentissage, nous conservons les annotations manuelles des segments d'obsolescence et nous projetons les annotations automatiques des indices linguistiques (outil ALIDIS, section 3.1). Cette annotation manuelle est la condition nécessaire pour mener une classification supervisée.

Disposant de méthodes d'apprentissage automatique éprouvées, nous considérons qu'elles permettent de quantifier la pertinence de nos indices dans le cadre de la classification supervisée. Cette tâche principale est au cœur de nos préoccupations pour évaluer l'intérêt de notre travail : si une méthode reconnue discrimine correctement les segments obsolescents des autres à l'aide de nos indices, elle montre leur intérêt prédictif pour le phénomène de l'obsolescence.

Nous évaluons les classifieurs en calculant un modèle (arbre, règles, etc.) sur 90 % du corpus d'apprentissage puis en l'appliquant et l'évaluant sur les 10 % qui restent. Cette validation croisée est répétée dix fois, de manière à ce que chaque segment textuel soit testé une fois et utilisé les neuf autres fois pour l'apprentissage. Les indicateurs fournis sont une moyenne sur les dix exécutions.

Nous avons effectué plusieurs expériences à l'aide de la plate-forme RAPID-MINER [14], qui permet d'utiliser des méthodes classiques de classification supervisée :

14. `http://www.rapidminer.com`

– arbres de décision (Quinlan, 1993) : l'entropie de chaque attribut par rapport aux classes permet d'en construire une hiérarchie dont les feuilles regroupent les instances d'une même classe ;

– bayes naïf (Hand et Yu, 2001) : utilisation d'un modèle probabiliste ;

– séparateurs à vaste marge (SVM - *support vector machine* (Boser *et al.*, 1992) :] une séparation linéaire optimale est réalisée en projetant dans un espace de dimension supérieure à l'aide d'un noyau non linéaire ;

– réseau de neurones (Minsky et Papert, 1969) :] calcul d'un hyperplan séparateur ;

– régression logistique (Kleinbaum, 1994) :] régression à partir d'un modèle logistique.

Les configurations livrées en standard par RAPIDMINER fournissent immédiatement de bonnes performances : les arbres employent un critère de gain ratio mais pas d'élagage, les SVM utilisent un noyau radial, etc. Ces résultats pourraient certainement être améliorés par un paramétrage adapté, mais requièrent une expertise que nous n'avons pas.

Enfin, nous avons utilisé des modèles à base de règles d'association (Agrawal *et al.*, 1993) : ce sont des expressions $X \rightarrow Y$ calculées à partir de motifs (ou combinaisons d'attributs) fréquents dans des contextes booléens. Les règles qui concluent sur un attribut de classe constituent un modèle pour la classification (Li *et al.*, 2001). Les modèles à base de règles d'association sont peu utilisés pour la classification des textes. Cependant, leurs performances sont comparables aux méthodes traditionnelles, les modèles fournis sont facilement interprétables, et nous possédons une expertise dans ce domaine (Rioult *et al.*, 2010). Les règles d'association méritent donc d'être testées sur ce problème.

Les résultats sont reportés au tableau 3. Ce tableau présente, pour chaque algorithme :

– la mesure d'aire sous la courbe ROC (*AUC, area under the receiver operating characteristic*) (Fawcett, 2003). Cette mesure indique la probabilité que le classifieur a de séparer les distributions des deux classes. Elle équivaut au test des rangs (Wilcoxon, 1945). L'intérêt de cette mesure est qu'elle n'est pas sensible au déséquilibre de population des classes ;

– pour chaque classe :

- la précision : proportion d'exemples correctement attribués à la classe,
- le rappel : proportion d'exemples de la classe retrouvés par le classifieur,
- le F-score, moyenne harmonique du rappel et de la précision.

De ces expériences, nous concluons que nos indices linguistiques sont pertinents pour discriminer les deux classes : toutes les méthodes obtiennent une aire sous la courbe ROC de 75 à 85 %. Ce bon comportement sur une mesure reconnue ne doit cependant pas ignorer ce qu'elle ne montre pas, à savoir les écarts de performances entre les différentes classes.

Algorithme	AUC	Classe non obsolescente			Classe obsolescente		
		Précision	Rappel	F-score	Précision	Rappel	F-score
Arbres	74,4	88,1	93,8	90,8	49,6	32,4	39,2
Bayes naïf	82,3	90,6	92,5	91,6	55,6	49,4	52,3
SVM	86,2	88,1	98,3	93,0	77,8	30,1	43,4
Réseau neurones	80,2	87,7	98,0	92,6	72,3	27,4	39,7
Régression logistique	80,9	92,2	82,6	87,1	40,6	63,0	49,4
Règles d'association	79,8	79,2	85,7	85,7	39,0	70,3	50,1

Tableau 3. *Performances (en pourcentage) pour la classification supervisée (validation croisée, 10-cross validation)*

Pour notre problème, seulement 15 % des segments sont obsolescents. Dans ce cas, un classifieur prônant systématiquement la classe majoritaire aurait un bon rappel (100 %) et une bonne précision (85 %) pour cette classe, mais aucun intérêt pour notre tâche. Il faut donc se concentrer sur les performances pour la classe obsolescente minoritaire (*cf.* les trois dernières colonnes du tableau 3).

Si certaines méthodes (SVM, neurones) obtiennent une bonne précision, là où d'autres (régression, règles d'association) ont un bon rappel sur la classe minoritaire, toutes les méthodes ont un F-score équivalent, de 40 à 50 %.

Nous privilégions dans la suite de cette présentation l'usage des règles d'association afin d'évaluer plus finement le pouvoir prédictif de nos indices linguistiques, car le modèle qu'elles définissent est aisément interprétable.

5. Apport des différents types d'indices linguistiques

Dans le but de préciser l'apport des différents indices linguistiques, nous effectuons maintenant les expérimentations suivantes :

– *(i)* comparaison à deux approches de bases, fondées sur des indices de surface élémentaires ou caractéristiques d'une expression temporelle ;

– *(ii)* apports des différents types d'indices (intraphrastiques, hiérarchiques, positionnels) ;

– *(iii)* évaluation par un expert sur un corpus de test différent du corpus d'apprentissage.

5.1. *Comparaison à des approches de base*

Le tableau 4 permet de mesurer les gains de notre méthode linguistique par rapport à deux systèmes base :

– base 1 : on ne prend en compte que la présence ou l'absence de chiffres (et donc par extension les expressions chiffrées et les dates) dans la phrase ;

– base 2 : on ne prend en compte que les indices les plus intuitifs : expressions temporelles déictiques, ponctuelles ou de durée lorsqu'elles réfèrent à une date proche du moment d'énonciation ou située dans le futur, les temps et modes futur et conditionnel, les adverbiaux exprimant un point de vue de type récence (*les territoires actuels*) ou prévision (*les recherches à venir*).

	Précision	Rappel	F-score
Base 1	23	31	26
Base 2	30	39	37
Tous les indices, en association	39	70,3	50,1

Tableau 4. *Comparaison des performances du classifieur par rapport à des systèmes de base*

Les résultats présentés dans ce tableau montrent le gain de notre méthode sur le rappel notamment. Ils montrent également qu'une technique trop basique est insuffisante pour traiter le problème posé.

5.2. *Évaluation*

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 analysons les variations de la performance en classification supervisée sur cinq vues différentes de notre corpus :

– corpusComplet : une vue qui prend en compte tous les indices ;

– corpusIPseuls : une vue qui prend en compte uniquement les indices intraphrastiques ;

– corpusIPHierar : une vue qui prend en compte les indices intraphrastiques et les indices hiérarchiques ;

– corpusIPPos : une vue qui prend en compte les indices intraphrastiques et les indices positionnels ;

– corpusEpure : un corpus épuré dans lequel sont enlevées les variables non significatives (en fonction des résultats des statistiques de base et de l'ACP).

Les résultats des performances de l'algorithme en précision et en rappel sur les différentes vues sont présentés dans le tableau 5.

	Précision	Rappel	F-score	AUC
corpusIPseuls	38	37	37,5	61,5
corpusIPHierar	39,9	45,6	42,5	66,2
corpusIPPos	33,2	56,7	41,9	68,6
corpusEpure	38,7	62,3	47,7	72,9
corpusComplet	39	70,3	50,1	79,8

Tableau 5. *Comparaison des performances du classifieur selon les différentes vues sur le corpus d'apprentissage*

Concernant les mesures de précision et de rappel (*cf.* tableau 5), les meilleurs résultats sont obtenus lorsque tous les types d'indices, linguistiques et discursifs, sont utilisés pour l'apprentissage des règles. On constate également que la valorisation des résultats des statistiques descriptives pour créer un corpus restreint aux seuls indices statistiquement pertinents apporte une nette plus-value.

Les résultats les moins bons sont obtenus lorsque seuls les indices intraphrastiques sont pris en compte. L'apport des indices hiérarchiques et positionnels est faible : la prise en compte des indices hiérarchiques favorise la précision alors que les indices positionnels privilégient le rappel.

6. Évaluation par les experts

Nous terminons cette partie expérimentale en relatant une application grandeur nature réalisée sur un corpus de test. Ce corpus de test est constitué de textes différents de ceux utilisés pour constituer le corpus d'apprentissage ; il n'est pas annoté manuellement et comprend environ 100 000 mots soit 3 916 phrases et titres. Les indices linguistiques (outil ALIDIS, *cf.* section 3.1) sont projetés sur ce corpus de test. Puis le modèle à base de règles d'association préalablement appris sur le corpus d'apprentissage est utilisé pour prédire le caractère obsolescent ou non des phrases du corpus de test.

Les résultats sont proposés aux experts afin qu'ils évaluent la classification automatique des phrases obsolescentes fournie par le classifieur. Les experts ont jugé l'ensemble des cas possibles :

– les phrases annotées obsolescentes par le classifieur et qui sont réellement obsolescentes ;

– les phrases annotées obsolescentes par le classifieur et qui ne le sont pas ;

– les phrases non annotées obsolescentes par le classifieur et qui sont réellement non obsolescentes ;

– les phrases non annotées obsolescentes par le classifieur et qui sont pourtant obsolescentes.

Nous avons comptabilisé les intersections des annotations automatiques avec la validation humaine des 3 810 phrases du corpus de test. Les résultats sont fournis par la matrice de confusion du tableau 6.

		Humain		
		Obsolescent	Non obsolescent	Total
Machine	Obsolescent	261	440	701
	Non obsolescent	138	2 971	3 109
	Total	399	3 411	3 810

Tableau 6. *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 l'on applique le coefficient 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 selon les accord entre deux juges, *cf.* section 2.3).

En termes de scores de pertinence, le tableau 7 rend compte de l'évaluation du classifieur automatique pour la classe *obsol* sur le corpus de test.

	Précision	Rappel	F-score
corpusComplet	0,37	0,65	0,47

Tableau 7. *Évaluation par les experts : performances pour la classe obsol*

En d'autres termes, l'outil automatique attire l'attention de l'expert sur le cinquième du corpus (701 sur 3 810 segments), contenant les deux tiers des phrases obsolescentes (261 sur 399 segments), un seul segment sur trois étant vraiment obsolescent (261 sur 701 segments). Dans l'absolu, ces résultats sont moyens : ils pourraient probablement être améliorés sur la base d'une définition plus précise de l'obsolescence et d'une annotation automatique des indices linguistiques plus fine. Cependant, dans un contexte professionnel, ce sont des résultats plutôt encourageants, sachant qu'aujourd'hui aucune mise à jour n'est effectuée de manière exhaustive dans les maisons d'édition [15] : on observe une nette réduction (de 3 810 à 701) du nombre de phrases à vérifier avec une précision médiocre mais un bon rappel.

Les variations de performances au cours de nos différentes expériences s'expliquent par des corpus aux provenances différentes : le corpus de test contient moins de textes appartenant au domaine de la géographie et plus de textes appartenant au domaine de l'histoire ; la proportion est inverse dans le corpus d'apprentissage. Il s'agit d'un problème délicat et qui confirme la forte dépendance de telles méthodes à la constitution des données d'apprentissage. Il faudrait idéalement un corpus d'apprentissage plus volumineux et plus homogène en termes de domaines (histoire, géographie,

15. La majorité des mises à jour est fonction de l'actualité et de décisions humaines (souvent politiques).

etc.) et donc aussi des annotateurs-rédacteurs disponibles pour réitérer le processus d'anotation manuelle, ce qui reste problématique dans un contexte professionnel.

7. Analyse des connaissances obtenues sur la base des associations émergentes

Les règles d'association présentent l'intérêt de fournir un modèle interprétable. Ainsi les règles apprises par le système puis utilisées pour le prototype nous renvoient des connaissances nouvelles en termes de combinaisons d'indices linguistiques. Par rapport à l'analyse en composantes principales, les connaissances apprises ici sont nettement plus locales et utiles que des tendances générales.

D'une manière générale, l'ensemble des types d'indices que nous avons décidé de prendre en compte est utile dans plusieurs règles : ceci montre l'intérêt de considérer des indices à granularité variable, dans un système où chacun des traits est nécessaire.

Cette section montre quelques exemples de combinaisons d'indices pertinentes pour l'obsolescence. En sortie du classifieur, les règles sont de la forme suivante :

$premierParag.position : debutDivision \wedge zone.rubriqueName : NULL \wedge title.entiteNom.classe : geopolitique \longrightarrow classe : obsol$

Cette règle stipule qu'une phrase qui contient une entité nommée de type géopolitique, qui est située dans un paragraphe en début de section et dont la rubrique textuelle est de type *NULL* (c'est-à-dire relevant d'une rubrique non référencée) sera considérée comme obsolescente par le classifieur.

7.1. *Des combinaisons d'indices comme marqueur de l'obsolescence*

L'association entre des indices hiérarchiques et des indices intraphrastiques est fréquemment caractéristique de l'obsolescence. 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

Nous observons également une forte attraction entre des indices positionnels textuels et des indices intraphrastiques : le premier paragraphe d'une division associée à une expression temporelle de type déictique coïncidence (*aujourd'hui*) ou à une

entité nommée de type géopolitique entraîneront souvent l'obsolescence du segment dans lequel l'indice intraphrastique 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 phrases au sein des paragraphes, les premières phrases de paragraphe contiennent souvent des indices temporels de type déictique coïncidence ou des entités nommées de type mesure évolutive lorsqu'elles sont obsolescentes. De plus, lorsqu'un verbe au conditionnel et une entité nommée de type géopolitique sont 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

L'obsolescence est également mise en valeur par l'association marquée entre plusieurs indices intraphrastiques. Ainsi, une phrase sera obsolescente si une expression temporelle de type déictique coïncidence est associée à une entité nommée de type géopolitique ou de type mesure évolutive.

§ 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

Des règles mettant en relation trois niveaux différents d'indices sont apprises par le système. Ainsi, 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 lieu référant à une ville, *(ii)* située en dernière position de paragraphe et *(iii)* chapeautée par un titre contenant une entité nommée de type géopolitique.

§ 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 3. *Combinaison de plusieurs indices intraphrastiques - 2*

Les résultats montrent enfin qu'un indice de type externe comme le domaine ou la rubrique joue un rôle important. Ainsi, une entrée de type géographie va être productive en phrases obsolescentes si la phrase contient une entité nommée de type mesure géopolitique, ce qui n'est pas le cas avec les textes « rubriqués » histoire si ce même indice est présent dans la phrase.

7.2. *Discussion*

Les règles apprises par le classifieur mis en place pour ce travail confirment nos intuitions sur la question du repérage automatique des segments d'obsolescence et, en même temps, s'intègrent aux résultats d'autres travaux de recherche connexes.

Ainsi, le point central sur lequel se basent nos traitements, à savoir la prise en considération d'indices de types différents et/ou de granularité variable, a été exploité par HoDac (2007) sur la question et le traitement de la position initiale, par Widlöcher (2008) sur les relations rhétoriques du discours ou encore par Bouffier (2008) sur le repérage de segments de recommandation dans des documents spécifiques (les *Guides de bonne pratique médicales*). Les travaux de Teufel (1999) mettent également en relation des indices de types aussi variés que ceux que nous exploitons dans le but de repérer automatiquement les phrases importantes (*argumentative zoning*) dans des écrits scientifiques à des fins de génération automatique de résumés d'article.

La prise en compte de tels indices discursifs, envisagés en termes de combinaisons, est particulièrement riche. L'obsolescence est un phénomène complexe qui ne peut se réduire à une liste simple associant un indice à une fonction. Au contraire, les indices linguistiques sont exploitables pour le repérage de l'obsolescence uniquement s'ils sont envisagés en termes de combinaisons ou de configurations. Bouffier (2008) insiste elle aussi particulièrement sur l'importance de la relation entre des indices lexicaux et des indices visuels dans les segments de type recommandation.

Concernant l'importance des marqueurs de type positionnel que nous mettons en évidence dans ce travail, nos résultats vont dans le sens des travaux de Marcu (2000) 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.

Les indices hiérarchiques, c'est-à-dire les indices présents dans un titre, sont des indices pertinents pour la recherche de segments obsolescents. Les traitements statistiques convergent vers des constats similaires : par exemple 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 (c'est-à-dire de connecteurs discursifs) et d'indice de nouveauté (c'est-à-dire de point de vue) dans les titres.

La caractérisation du texte en fonction des rubriques thématiques est enfin une information centrale : que ce soit à travers les statistiques descriptives ou l'apprentissage automatique, 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 de lieu + entité nommée de 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).

8. Conclusion et perspectives

Traiter l'obsolescence pour aider à la mise à jour des textes est une démarche originale. Nous proposons une méthodologie permettant à la fois de la caractériser et de la comprendre mais également de la repérer de manière automatique.

L'obsolescence est un phénomène réel (entre 10 et 15 % du nombre de phrases de notre corpus) et relativement flou (le consensus entre les annotateurs n'est pas total). Les indices sémantiques et discursifs que nous repérons et annotons de manière automatique se révèlent être de bons indices pour répondre à notre objectif de recherche de segments obsolescents.

Nous avons montré à l'aide d'une large variété de modèles (arbres, SVM, règles) que nos indices sont pertinents pour une tâche de classification supervisée et permettent de réduire la tâche manuelle de repérage.

En particulier, les analyses effectuées sur les modèles à base de règles d'association nous encouragent dans l'idée que les combinaisons d'indices linguistiques (sémantiques et discursifs) sont pertinentes pour le repérage automatique de l'obsolescence : d'un côté ces règles s'inscrivent dans la lignée des travaux en linguistique et en TAL, de l'autre, les scores fournis valident par ailleurs la pertinence de nos indices. Parallèlement à ce constat, ces résultats confirment nos choix méthodologiques à considérer des indices de types différents, de niveaux de granularité variés, du mot au discours.

La mise en œuvre d'une double évaluation du prototype d'aide à la mise à jour a permis :

– une évaluation intrinsèque du système par validation croisée qui a produit des résultats intéressants ;

– une évaluation par les experts, qui confirme les résultats précédents.

Ces évaluations fournissent, par ailleurs, une estimation *a posteriori* de la qualité de notre outil d'annotation automatique des indices linguistiques (*cf.* outil ALIDIS) et sur les possibilités de l'améliorer. 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 plus 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.

Ce travail ouvre plusieurs perspectives. Tout d'abord, il nous semble important d'affiner les indices, de les regrouper par *valeurs* sémantiques (par exemple créer un indice *déictique* qui regrouperait des adverbiaux de type *ponctuel* et *inachevé*) dans le but de rendre le modèle plus pertinent et plus efficace. De plus, il serait pertinent de différencier les types de mises à jour selon les types d'obsolescence : une vision plus précise du phénomène de l'obsolescence serait possible et la prise en compte des besoins de l'utilisateur serait meilleure. Enfin, une caractérisation plus précise des règles selon les rubriques thématiques (géographie, économie, histoire, etc.) est nécessaire : cela permettrait d'affiner les résultats encourageants obtenus sur des corpus hétérogènes.

9. Bibliographie

Agrawal R., Imielinski T., Swami A., « Mining association rules between sets of items in large databases », *Proc. of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, USA*, p. 207-216, 1993.

Biber D., « A typology of english texts », *Linguistics*, vol. 27, p. 3-43, 1989.

Boser B. E., Guyon I. M., Vapnik V. N., « A training algorithm for optimal margin classifiers », *5th Annual ACM Workshop on COLT*, p. 144-152, 1992.

Bouffier A., Analyse discursive automatique de textes - Application à la modélisation de textes incitatifs, Thèse de doctorat, Université Paris Nord - Villetaneuse, 2008.

Charolles M., « L'Encadrement du Discours, Univers, Champs, Domaine et Espaces », *Cahiers de Recherche linguistique*, 1997.

Edmondson W. J., « If coherence is achieved, then where doth meaning lie ? », *in* W. Bublitz, U. Lenk, E. Ventola (eds), *Coherence in spoken and written discourse, how to create it and how to describe it*, John Bejamins, p. 251-265, 1997.

Fawcett T., ROC Graphs : Notes and Practical Considerations for Researchers, Technical report, HP Laboratories, 2003.

Gosselin L., *Temporalité et modalité*, de Boeck.Duculot, 2005.

Grosz J., Sidner A., « Attention, intentions, and the structure of discourse », *Computational linguistics*, july-sept, 1986.

Hand D., Yu K., « Idiot's Bayes - not so stupid after all ? », *International Statistical Review*, vol. 69, n° 3, p. 385-399, 2001.

Ho-Dac M., Jacques M.-P., Rebeyrolles J., « Sur la fonction discursive des titres », *in* S. Porhiel, D. Klingler (eds), *L'unité texte, Actes du colloque Regards croisés sur l'unité texte / Conjoint Perspectives on Text*, Actes des 4èmes Journées de Linguistique de Corpus, Perspectives, Chypre, 2004.

HoDac M., La position initiale dans l'organisation du discours : une exploration en corpus, Thèse de doctorat, Université de Toulouse 2 - Le Mirail, 2007.

Hripcsak G., Heitjan D., « Measuring agreement in medical informatics reliability studies », *Journal of Biomedical Informatics*, vol. 35, n° 2, p. 99-110, 2002.

Ibekwe-SanJuan F., « Annotation d'indices de nouveautés dans les écrits scientifiques et techniques », *Colloque Indice, Index, Indexation*, 2005.

Kleinbaum D., *Logistic regression. A self-learning text*, Springer-Verlag, 1994.

Laignelet M., Analyse discursive pour le repérage automatique de segments obsolescents dans les documents encyclopédiques, Thèse de doctorat, Université de Toulouse - Le Mirail, 2009.

Laignelet M., Rioult F., « Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs », *Actes de TALN 2009*, 2009. prix du "Best Paper".

Lebart L., Morineau A., Piron M., *Statistique exploratoire multidimensionnelle*, Dunod Paris, 1995.

Li W., Han J., Pei J., « CMAR : Accurate and Efficient Classification Based on Multiple Class-Association Rules », *IEEE International Conference on Data Mining*, 2001.

Marcu D., « The Rhetorical Parsing of Unrestricted Texts : A Surface-Based Approach », *Computational Linguistics*, 2000.

Minsky M. L., Papert S. A., *Perceptrons*, MIT Press, 1969.

Quinlan J. R., *C4.5 : Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

Reichenbach, *Elements of symbolic logic New- york*, Free-Press,, 1966.

Rioult F., Zanuttini B., Crémilleux B., *Advances in Intelligent Information Systems*, vol. 265 of *Studies in Computational Intelligence*, Springer, chapter Nonredundant generalized rules and their impact in classification, p. 3-25, 2010.

Sebastiani F., « Machine learning in automated text categorization », *ACM Computing Surveys*, vol. 34, n° 1, p. 1-47, 2002.

Teufel S., Argumentative Zoning, PhD thesis, Université de Edimbourg, 1999.

Widlöcher A., Analyse macro-sémantique des structures rhétoriques du discours. Cadre théorique et modèle opératoire., Thèse de doctorat, Université de Caen, 2008.

Widlocher A., Bilhaut F., « La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus », *Actes de la 12e Conference TALN*, Dourdan, France, 2005.

Wilcoxon F., « Individual comparisons by ranking methods », *Biometrics*, 1945.

Zerida N., Lucas N., Crémilleux B., « Combinaison de descripteurs linguistiques et de structure pour la fouille d'articles biomédicaux », *Schedae*, p. 69-78, 2006.