Traitement automatique d'informations textuelles complexes : connaissances linguistiques hétérogènes et à granularité variable

Marion Laignelet

Université Toulouse 2, Laboratoire CLLE-ERSS
5 allée Antonio Machado, 31058 Toulouse Cedex 9
marion.laignelet@univ-tlse2.fr

Résumé. Dans cet article, nous présentons une méthodologie permettant le traitement et la structuration de données linguistiques complexes. Par données complexes, nous envisageons des informations textuelles présentant la particularité d'être à la fois hétérogènes sémantiquement et à granularité variable. Pour passer d'une structure linguistique constituée d'objets complexes à une organisation des données permettant l'application de méthodes statistiques et/ou de fouille de données, nous proposons un modèle de représentation des unités du discours. Ce travail est mené dans le cadre d'un projet visant la mise en oeuvre d'un prototype d'aide à la mise à jour de documents encyclopédiques articulé autour du repérage automatique de zones textuelles contenant de l'information obsolète.

1 Introduction

Dans cet article, nous proposons un modèle de représentation des unités discursives qui nous permet de passer d'annotations linguistiques complexes à une représentation des informations manipulables par des outils statistiques et/ou de fouille de données.

En linguistique, les objets étudiés sont divers et relativement stables au sein des communautés : le morphème pour les morphologues, le mot, lexie ou lexème pour le lexicologue, les catégories syntaxiques pour les syntacticiens (Brunet, 2003; Kao et Poteet, 2007). L'utilisation des statistiques en linguistique n'est pas nouvelle : les morphologues utilisent les statistiques pour prédire les schémas de formation des mots (Grabar et Zweigenbaum, 1999; Daille et al., 2001; Jacquemin, 1997; Lavallée et Langlais, 2010), les lexicologues et sémanticiens étudient les fréquences d'apparition de termes en collocation (Bouillon et al., 2000; Hearst, 1992; Sébillot et al., 2000) ou la recherche de motifs (Nouvel et al., 2010), les syntacticiens quant à eux s'intéressent aux structures récurrentes à l'intérieur des phrases (Brill, 1992; Schmid, 1994; Dejean et al., 2010).

D'un point de vue applicatif, ce travail a été élaboré dans le cadre d'un projet de recherche visant le repérage automatique de segments textuels contenant de l'information potentiellement obsolète. Repérer de tels segments permet *in fine* d'aider des rédacteurs à mettre à jour les contenus d'encyclopédies.

Mais l'obsolescence n'est pas un phénomène qui se laisse appréhender à l'aide d'outils simples. Il n'existe pas de "lexique de l'obsolescence" ni de catégorie grammaticale qui permette de le repérer facilement. Au contraire, il s'agit d'un phénomène complexe qui, pour pouvoir être repéré de manière automatique, doit faire appel à des informations linguistiques hétérogènes d'un point de vue sémantique et variées d'un point de vue du grain d'analyse. Cela implique donc qu'il n'est pas possible de se limiter à un niveau linguistique spécifique mais qu'au contraire il est nécessaire d'exploiter conjointement tous les niveaux : mot, groupes de mots, syntagmes, paragraphes, titres, genres. Ce travail se situe dans le cadre des études sur le discours où la RST ou la SDRT sont prépondérantes.

La Rhetorical Structure Theory (Mann et Thompson, 1988) décrit la cohérence des textes à travers l'identification de relations rhétoriques reliant des segments de discours. La Segmented Discourse Representation Theory (Busquets et al., 2001) est également une théorie de l'organisation des textes basée sur l'idée que des segments (*i.e.* des unités élémentaires) sont reliés entre eux par des relations de cohérence (*i.e.* rhétoriques). Des travaux en cours (Péry-Woodley et al., 2009) exploitent la notion de marqueur pour rendre compte de manière automatique de la structure des discours.

Nous nous distancions de ces théories, notamment parce que l'obsolescence n'est pas une catégorie du discours qui y est étudiée. En revanche, notre objet d'étude est le même : le discours. Nous défendons l'idée qu'il s'agit d'une unité linguistique valide composée d'autres unités linguistiques hétérogènes et à granularité variable : le mot, le sytgme le paragraphe, la position sont porteurs de sens et c'est ensemble qu'il faut les exploiter. Ainsi, ce travail s'inscrit à la suite des travaux de Biber et al. (2007) qui propose une méthodologie inductive pour la classification automatique des textes basé sur la prise en compte d'indices linguistiques variés. Les travaux de Teufel et al. (1999); Teufel (1998) montrent qu'il est esssentiel de prendre en compte des indices à granularité variable pour mettre en lumière les principes argumentatifs dans les textes. Enfin, Mani (2001) a montré que dans le cadre d'une tâche de résumé automatique, la prise en compte de la position des éléments dans le texte est importante : une phrase sera jugée importante si elle se situe dans l'introduction ou dans la conclusion du document à résumer.

Dans une première section, nous rappelons l'arrière plan théorique qui guide nos travaux. Les questions de l'unité d'analyse, de l'objet d'étude et de la nécessité de s'inscrire dans un modèle adaptable sont notamment présentés. La section 3 décrit le phénomène de l'obsolescence ainsi que le corpus mis en oeuvre. Dans la section 4, nous présentons les différents traits linguistiques repérés de manière automatique dans les textes. Nous insistons notamment sur leur caractère hétérogène et variable. Nous décrivons ensuite, dans la section 5, notre proposition de modèle de représentation de ces données linguistiques qui permet de passer d'un mode de représentation textuel à un mode de représentation conceptuel. Enfin, nous illustrons notre démarche à travers un exemple d'application : la recherche des combinaisons de traits linguistiques pertinentes, à l'aide d'outils statistiques, pour déterminer si un segment contient ou non de l'information obsolète et donc s'il convient de la marquer comme étant à mettre à jour.

2 La question de l'unité d'analyse dans les textes

Quand on parle de statistiques sur des données linguistiques, l'objet étudié peut concerner les graphies, les n-grammes, les lemmes, les classes de fréquence, les codes grammaticaux,

ou encore les structures syntaxiques et les structures sémantiques. La plupart du temps, un seul niveau d'analyse linguistique est pris en compte (Kao et Poteet, 2007). Or un texte n'est pas un sac de mot, encore moins un sac de fréquences (Péry-Woodley, 2005). Avant de se demander quel test statistique il convient d'appliquer pour une tâche donnée, la question de l'unité d'analyse est centrale.

2.1 Qu'est ce qu'un texte ?

Un texte est un objet structuré, organisé, hiérarchisé au sein duquel les mots entretiennent des relations particulières entre eux mais également avec les phrases, les paragraphes ou encore les titres. Un texte est constitué de segments textuels de granularité plus ou moins locale, plus ou moins globale qui s'organisent entre eux selon divers modes.

D'une manière générale, l'organisation des textes peut être vu, comme le définit la théorie de la compositionnalité holiste, selon 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 ces deux principes 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. Nous cherchons à comprendre comment des éléments locaux s'organisent ensemble, dans des structures textuelles diverses, pour permettre l'interprétation de l'information.

Le terme de *discours* est dans ce cadre préféré à celui de texte. Le 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. En effet, dans notre approche, nous prenons en compte des 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 parce que tous ces éléments participent de la construction finale du sens du discours.

Malgré l'apparente linéarité textuelle, le texte est un objet complexe et structuré au sein duquel différents mécanismes textuels et discursifs entrent en jeu. Les relations et hiérarchies entre les segments textuels sont en partie expliquées à travers les notions de cohérence et de cohésion (Johnson et Johnson, 1999, p. 55), (Carter-Thomas, 2001, p. 35), (Degand et Sanders, 2002).

Dans *Cohesion in English*, Halliday et Hasan (1976) font 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 donc la manière dont se construit un texte sur la base des relations de signification entre ses éléments.

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. Mais rechercher les indices, les marques linguistiques qui font sens nous semble une piste intéressante.

2.2 Traits linguistiques et marqueurs de discours

Un discours est donc un tout unifié, une unité qui tire son sens de l'existence et de la relation entre ses parties, lesquelles parties s'avèrent être hétérogènes et à granularité variable. Pour repérer de manière automatique ces parties et les relations les unissent, nous nous inscrivons dans une approche de TAL visant l'identification de marqueurs de discours.

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 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.

Péry-Woodley (2000) définit les marqueurs discursifs comme des *« traces qui constituent une signalisation orientant l'interprétation »*. Les connecteurs (« Et », « mais », « pour résumer », « en somme ») sont des marqueurs discursifs. Les *marqueurs d'intégration linéaires* (MIL, Jackiewicz (2005)) (« Premièrement », « D'une part [...], d'autre part ») participent également de l'organisation textuelle.

Selon Degand et Sanders (2002), pour signaler au lecteur la cohérence globale du texte, le scripteur dispose de *global discourse markers* (GDM). Ce sont des expressions dont le rôle est pragmatique et dont la fonction est organisationnelle [1]. Les GDM sont de trois types :

- les *metadiscourse markers* réfèrent explicitement à l'organisation du discours : partitionnement (« first », « then », etc.), étapes clés du texte (« in sum », « to conclude », etc.), buts discursifs ou de l'énonciateur (« I will argue », « My purpose is », etc.), changements 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 *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 changements thématiques.

Dans la classification proposée par HoDac (2007), les indices linguistiques sont organisés relativement à leur fonction (textuelle, idéationnelle ou interpersonnelle). Le caractère hétérogène et à granularité variable des types d'indices y est également mis en valeur.

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

– les *indices textuels* peuvent être linguistiques et comprennent les marqueurs méta-discursifs, les anaphores, etc. ; ils peuvent être typographiques (surlignements ou mises en gras) ; enfin ils peuvent être propres à la structure du texte, comme les titres.
– les *indices texto-idéationnels* apportent un sens instructionnel et un sens propositionnel (les syntagmes prépositionnels dans le cadre de l'hypothèse de l'encadrement du discours, (Charolles, 1997).
– les *indices texto-interpersonnels* : il s'agit par exemple des adverbes modalisateurs, des constructions impersonnelles.

Ces différentes approches théoriques sont basées sur la notion de marqueur. Dans ce travail, nous considérons que tout objet linguistique est potentiellement un marqueur : tant qu'aucune fonction particulière ne lui est associée, il est un trait linguistique. Il ne devient marqueur que lorsqu'une relation peutr être établie entre lui et une fonction du discours spécifique. Nous insistons par ailleurs sur le fait qu'un marqueur peut se réaliser soit par un trait linguistique seul, soit par une configuration de traits.

Pour conclure, nous entendons par données complexes des données linguistiques (ou traits) hétérogènes d'un point de vue sémantique et à granularité variable. La difficulté réside alors dans la manipulation de telles données par des outils statistiques qui nécessite un aplanissement et une uniformisation des données discursives brutes, *i.e.* telles qu'elle apparaissent dans les textes.

Avant de dresser un panorama des traits linguistiques que nous exploitons, la section suivante présente le phénomène que nous souhaitons repérer de manière automatique, l'obsolescence.

3 L'obsolescence

La recherche des segments d'obsolescence, c'est-à-dire des segments textuels qui nécessitent potentiellement une mise à jour de leur contenu, est une tâche complexe. L'obsolescence ne se réduit pas à un marqueur rhétorique spécifique mais semble au contraire nécessiter la prise en compte de nombreux indices linguistiques tels que ceux que nous venons de présenter.

3.1 Les segments obsolescents

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(s), de modification(s) ou de changement(s) 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 à mettre à jour. L'auteur décrit l'état des recherches actuelles sur le Sida.

Dans cet extrait, toutes les phrases méritent une vérification : il y a donc autant de segments d'obsolescence qu'il y a de phrases. L'unité phrastique est prise en compte pour deux raisons principales : d'abord parce qu'elle fait sens pour tout le monde et notamment pour les

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*

rédacteurs experts des éditions Larousse avec qui nous avons travaillé ; puis, techniquement, parce que c'est une unité plus facile à repérer de manière automatique que la proposition par exemple (sur la base de la prise en compte de la ponctuation mais aussi à l'aide de lexiques de sigles).

3.2 Description du corpus et annotation manuelle

Le corpus sur lequel nous travaillons est composé d'articles extraits des encyclopédies suivantes : l'encyclopédie Mémofiches des éditions Atlas ainsi que deux encyclopédies des éditions Larousse (le Grand Larousse Informatisé et le Grand Universel Larousse). Ce corpus est constitué de 282 141 mots. Il contient 9 916 phrases et 1 711 titres.

Les phrases ont été annotées manuellement selon leur caractère obsolescent ou non : un expert linguiste [2] a annoté le sous-corpus ATLAS et le sous-corpus LAROUSSE, lui-même également annoté par trois experts rédacteurs [3]. La part de segments obsolescents s'élève à 15 % des phrases (1508 segments sur 9916).

Parce que nous disposons d'une multi-annotation humaine [4], 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 cœfficient 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 cœfficient r de Finn [5] (Hripcsak et Heitjan, 2002). Ce cœfficient permet d'aplanir la dispro-

2. Moi-même, Marion Laignelet.
3. Rédacteurs de la société Larousse.
4. Le sous-corpus LAROUSSE a été annoté par quatre experts différents.
5. Nous utilisons l'algorithme existant dans le logiciel R.

portion 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 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).

Ces chiffres 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.

Le corpus ENCYCLO est donc annoté manuellement des segments d'obsolescence. Il reçoit dans un second temps l'annotation automatique des traits linguistiques de l'outil ALIDIS que nous présentons dans la section suivante.

4 Les traits linguistiques dans les textes

Cette section précise la démarche linguistique employée pour repérer de façon automatique des traits linguistiques dans les textes. Ils ont été définis relativement à notre objectif de recherche des segments d'obsolescence mais ils sont suffisamment génériques pour être décrits de façon indépendante.

Leur repérage est intégré dans un processus automatisé développé à l'aide de la plateforme LinguaStream [6] (Widlöcher et Bilhaut, 2005). Cette plateforme de développement de TAL (traitement automatique des langues) facilite la création de chaînes de traitements complexes sur la langue : segmenteurs, lexiques, grammaires ProLog, Macro Expressions régulières, etc. L'outil ALiDis [7], développé à l'aide de cette plateforme (Laignelet, 2009), regroupe différents modules de repérage et d'annotation sémantique automatique : traitement du temps, repérage des entités nommées, annotation du point de vue de l'auteur, etc.

4.1 Panorama des traits textuels

En plus de leur diversité au niveau sémantique, ces traits sont multi-échelle, c'est-à-dire qu'ils apparaissent à différents niveaux textuels.

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

Une première classe de traits concerne les **informations temporelles**. L'analyseur temporel mis en œuvre repère et annote sémantiquement les 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 modèle temporel ne rend pas compte de manière exhaustive du découpage temporel du texte, de l'ancrage des événements dans une référence temporelle ou encore de la succession des

6. http ://www.linguastream.org/whitepaper.html

7. Pour ANNOTATION LINGUISTIQUE DES DISCOURS, disponible sur `http://marion.laignelet.free.fr/ressources`

événements sur la ligne du temps (Reichenbach, 1947; Gosselin, 2005). Nous exploitons deux types d'information.

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 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*. Pour chacune des expressions temporelles repérées, nous calculons les cinq valeurs suivantes, identifiées relativement aux besoins en termes de pratique éditoriale : *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.

Voici quelques exemples :

(1) pendant dix ans $\begin{bmatrix} nature : durée \\ sitTps : indéterminé \end{bmatrix}$

(2) en 1930 $\begin{bmatrix} nature : ponctuel \\ sitTps : antériorité++ \end{bmatrix}$

(3) depuis 2005 $\begin{bmatrix} nature : inachevée \\ sitTps : coïncidence \end{bmatrix}$

(4) aujourd'hui $\begin{bmatrix} nature : déictique \\ sitTps : coïncidence \end{bmatrix}$

Le typage des temps verbaux est directement issu des résultats de l'analyseur morpho-syntaxique TreeTagger [8] disponible comme module externe dans LinguaStream. Le typage des verbes est le suivant : le trait `temps` peut prendre les valeurs suivantes : *passé-composé, passé antérieur, plus-que-parfait, futur antérieur, conditionnel passé, présent, passé simple, imparfait, futur, conditionnel*.

Nous prenons également en compte des **informations aspectuelles et modales** à travers, entre autres, le repérage de périphrases verbales. 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. L'exemple suivant illustre un type d'expression repérée et annotée par cet analyseur :

(5) Les 3e et 4e tranches [...] [sont en cours de] $\begin{bmatrix} accomplissement : déroulement \end{bmatrix}$ réalisation.

Un analyseur d'entités nommées repère et type des expressions variées. Cet analyseur est principalement basé sur l'utilisation de lexiques et de garmmaires locales. Nous nous intéressons à huit classes, certaines étant susceptibles de se subdiviser en sous-classes. Ainsi une expression classée comme un *lieu* peut se subdiviser en sous-classes de type *rivière, pays, ville, montagne*. Par exemple :

8. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

(6) Les expéditions d'Arzila [à Tanger] $\left[\begin{array}{l} classe : lieu \\ sousClasse : ville \end{array}\right]$

Cet analyseur repère également des expressions relevant des classes *personne*, *sigle*, *web*, *mail*, *marque*, *géopolitique* ou encore *mesure*.

(7) La majeure partie du butin ramené par [Drake] $\left[\, classe : personne \,\right]$

(8) Le trafic de la [VOC] $\left[\, classe : sigle \,\right]$ est centré sur le commerce du poivre et d'autres épices.

(9) les litiges supérieurs à [7 600 euros] $\left[\begin{array}{l} classe : mesure \\ sousClasse : évolutif \end{array}\right]$

(10) Le [taux de chômage] $\left[\, classe : géopolitique \,\right]$ s'est effectivement effondré.

Une dernière classe de traits concerne **les expressions exprimant un point de vue**. L'analyseur d'expressions du point de vue donne des informations sur le positionnement de l'auteur vis-à-vis des propos qu'il écrit. Nous proposons neuf types différents : *distance, jugement, récence, prévision, importance, jugement personnel, source, thématique, restriction, argumentation, superlatif*.

Le type *jugement* donne des informations modalisantes sur le point de vue de l'auteur :

(11) 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.

Les types *récence* et *prévision* sont orientés temporellement. Il s'agit de syntagmes nominaux contenant un adjectif temporel qui ne peut être interprété que si l'on connait la date à laquelle l'article a été écrit :

(12) le [dernier Mondial] $\left[\, type : récence \,\right]$ s'est tenu à

(13) les [recherches à venir] $\left[\, type : prévision \,\right]$

Le rédacteur peut également insister sur l'importance d'un fait à un moment donné :

(14) Il s'agit d'[un véritable enjeu] $\left[\, type : importance \,\right]$

Il peut également se distancier des propos qu'il avance en citant ses sources :

(15) on estime $\left[\, type : distance \,\right]$

(16) Selon le rapport de l'INSEE $\left[\, type : source \,\right]$

L'introduction de définitions introduit une certaine volonté de clarification des propos :

(17) On distingue $\left[\, type : définition \,\right]$ deux classes :

Enfin, nous prenons en compte les organisateurs de l'argumentation car ils permettent au rédacteur de structurer ses propos. Ainsi, nous repérons des éléments argumentatifs comme « d'abord », « puis », « dans un premier temps », ainsi que des expressions comme « À ce sujet/propos » ou encore « Pour ce qui est de la dette ».

L'ensemble de ces traits syntagmatiques peuvent être réalisés au sein de l'unité phrase : dans ce cas, nous parlons de **traits intra-phrastiques**. Ils peuvent également apparaître dans des titres : nous parlons alors de **trait hiérarchique**. Les traits hiérarchiques correspondent au

fait qu'un élément peut être sous la dépendance d'un autre, comme dans la relation existant entre une phrase et un titre. Une des conséquences est que les types de traits repérés dans les phrases sont également repérés dans les titres.

Parallèlement aux traits intra-phrastiques et hiérarchiques, nous traitons **les indices positionnels**. Ils peuvent être de deux types. Les traits positionnels phrastiques rendent compte de la position des traits intra-phrastiques au sein d'une unité (début et fin de phrase par exemple ou encore première phrase ou dernière phrase d'un paragraphe). Le niveau de hiérarchie dans le document, principalement rendu par les titres, est aussi exploité.

Les traits 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*).

4.2 Performance de l'outil d'annotation des traits linguistiques

La performance des modules de repérage automatique a été évaluée manuellement sur un corpus d'environ 50 000 mots. Les résultats sont donnés dans le tableau suivant : pour chaque type de trait est indiquée la précision (proportion d'indices correctement retrouvés) et le rappel (proportion d'indices retrouvés).

	Nombre d'occurrences	précision	rappel
Temps Verbaux	15 768	97 %	98 %
Adverbiaux Temporels	4 459	92 %	98 %
Périphrases Verbales	85	99 %	43 %
Entités Nommées	12 306	99 %	83 %
Expression du Point de Vue	916	73 %	98 %
Moyenne	33 534	93 %	85 %

TAB. 1 – *Performance globale de l'outil* ALIDIS

On observe quelques disparités de performance selon les analyseurs. 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 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 « la loi [prévoit] un an de prison pour... » ; or le fait que nous nous basons sur la présence du verbe *prévoir* nous renvoie les deux cas.

D'une manière générale, ces résultats sont suffisants pour envisager un traitement statistique à grande échelle, du moins dans un premier temps. Il sera probablement nécessaire terme de pallier ces disparités en pondérant chacun des traits ou classes de traits en fonction de leur fiabilité.

Dans la section suivante, nous proposons un exemple d'annotation automatique du corpus appliqué sur un extrait issu du corpus.

4.3 Un exemple d'annotation automatique

Les figures 3 et 4 montrent comment et avec quels indices l'exemple 2 est annoté. La première figure illustre le cas des indices intra-phrastiques ; la seconde, le cas des indices discursifs (hiérarchiques et positionnels).

x. 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.

Source : Corpus ATLAS (fiche Médecine - Le Sida)

Exemple 2 – *Un segment d'obsolescence*

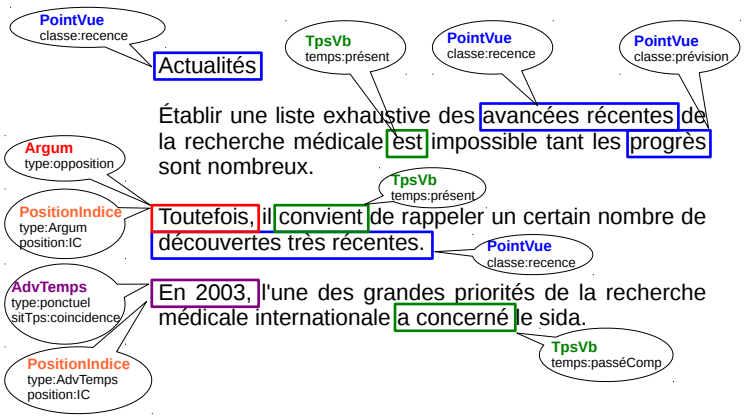

FIG. 3 – *Repérage et annotation automatique des indices intra-phrastiques*

Le corpus se voit ainsi enrichi d'une annotation automatique de traits linguistiques variés en termes de type et de granularité. Les objets linguistiques que nous souhaitons traiter à l'aide d'outils statistiques sont hétérogènes (tailles différentes, recouvrements des objets, etc). Il s'avère alors nécessaire de transformer les repérages et annotations en un format exploitable par des outils statistiques. Pour cela, nous proposons une étape intermédiaire de modélisation des données.

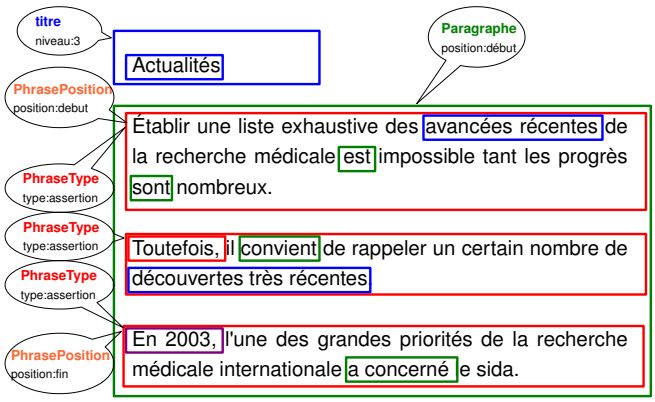

FIG. 4 – *Repérage et annotation automatique des indices discursifs : titres, phrases et positions*

5 Un modèle de représentation des indices de discours

Pour pouvoir traiter statistiquement l'ensemble de nos données, il nous faut passer d'un format textuel à une matrice (soit un tableau d'individus caractérisés par un ensemble de variables). La principale difficulté concerne la gestion des différences de granularité des traits linguistiques.

Pour arriver à cette transformation, nous proposons un modèle de représentation des traits textuels permettant de rendre compte de leur organisation. Il est conçu pour suivre les quatre principes de représentation présentés dans le tableau 2.

Principe 1	Un titre et une phrase doivent pouvoir être décrits à travers n'importe quel type de trait : intra-phrastique, hiérarchique, positionnel phrastique et textuel, externe
Principe 2	Une phrase peut être obsolescente mais un titre ne le sera jamais, il est un prédicteur potentiel de l'obsolescence (*cf.* la notion d'héritage de contexte Zerida et al. (2006))
Principe 3	On ne connait pas *a priori* le nombre de traits présents à l'intérieur de la phrase et du titre ni le niveau de profondeur du segment ; de plus, on veut pouvoir modifier les traits et leur typage sans avoir à réécrire tous les programmes.
Principe 4	On ne connait pas *a priori* le format d'entrée pour les statistiques ; le stockage des données doit être au plus près de la réalité des textes (en termes de relations entre les traits et les segments que l'on souhaite décrire).

TAB. 2 – *Les principes de représentation*

Le principe 2 découle du fait que dans notre corpus, les titres ne nécessitent pas de mise à jour[9]. Par ailleurs, nous craignons une redondance informationnelle si les traits textuels présents dans les titres sont traités à la fois comme segments obsolescents et aussi comme segments prédicteurs de l'obsolescence. Cela fausserait sans doute les résultats.

Le principe 3 permet de rendre l'outil évolutif : nous souhaitons un système facilitant l'ajout, la suppression ou la modification des traits textuels et/ou de leur typage sans avoir à tout réécrire à chaque modification. Le nombre et la nature des traits (variables) décrivant chaque phrase est donc calculé dynamiquement en fonction des données d'entrée.

Les objets linguistiques et les relations qu'ils observent entre eux sont représentés dans le schéma UML de la figure 5.

Dans ce schéma UML les unités discursives sont définies comme des unités élémentaires pour nos besoins ainsi que les relations qu'elle peuvent entretenir entre elles. Tout élément linguistique annoté est considéré comme une unité discursive (et donc comme un trait linguistique). À chaque unité discursive est associée une structure de traits (apportant des informations de type sémantique, syntaxique, morphologique, etc.). Cette unité discursive peut être réalisée par n'importe quel élément linguistique qui a été annoté soit par l'outil ALIDIS, soit manuellement.

Parce qu'on cherche à traiter les relations entre certaines unités spécifiques, nous proposons ensuite de distinguer deux sortes d'unités discursives : les unités d'analyse et les indices discursifs.

Le choix des unités d'analyse est fait par le linguistique analyste. Dans notre cas, nous considérons les phrases et les titres. Mais parce que ces unités ont des rôles différents dans les textes, nous nous sommes donné la possibilité de les traiter de manière différente : certaines

9. Aucun titre n'a été annoté manuellement par les experts.

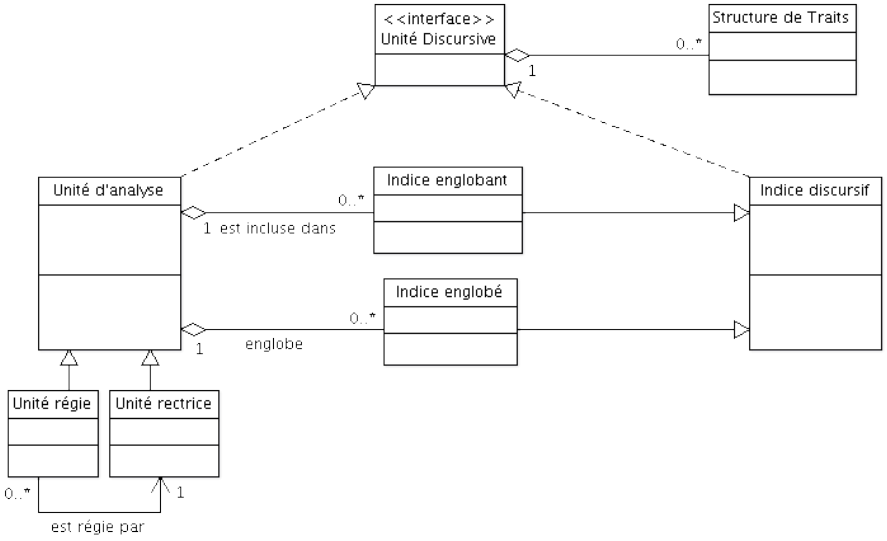

FIG. 5 – *Modèle UML de représentation des indices de discours : gérer la variabilité du grain d'analyse*

unités sont régies, d'autres sont rectrices. Ce choix permet de traiter les phrases comme des unités régies par des titres, et qui auront donc comme caractéristique d'hériter des traits des unités rectrices, les titres en l'occurrence.

Les indices discursifs sont réalisés par les autres unités existantes dans le texte. Ils peuvent être de deux types. Les indices englobés représentent les indices inclus dans l'unité d'analyse en question : dans notre cas, il s'agit des éléments intra-phrastiques essentiellement. Les indices englobants sont les indices qui sont de taille plus grande que l'unité d'analyse : dans notre cas, il s'agit des éléments positionnels essentiellement. Une unité d'analyse peut être associée à *n* indices englobants et *n* indices englobés.

Les limites principales de ce modèle sont :

– la non-représentation des relations linéaires entre unités de même classe (notamment entre indices englobés) ;
– la position des expressions linguistiques dans la phrase est aujourd'hui mal gérée : la position est un indice qui s'ajoute à l'expression linguistique (elle aussi un indice) : il y a donc redondance et, de fait, un biais interprétatif.

Cette étape de modélisation met sur le même plane l'ensemble des éléments linguistiques tout en conservant les informations sémantiques et syntaxiques propres à chacun (à l'aide des structures de traits). Cette étape est la base de la transformation et de l'organisation à la fois

des annotations manuelles de l'obsolescence des segments (*cf.* section 3.2) et des annotations automatiques des traits textuels (*cf.* outil ALIDIS, section 4.2) dans une base de données relationnelle (Laignelet, 2009). L'association des deux annotations (manuelle et automatique) permet la mise en place d'un système d'apprentissage supervisé (*cf.* section 6).

Nous disposons en fin de traitement alors pour notre corpus de 9916 individus (initialement les phrases du corpus) qui peuvent être décrites relativement à 150 variables différentes (initialement les indices linguistiques, tous types et tous grains confondus, repérés automatiquement). Chaque individu est également caractérisé selon sa nature obsolescente ou non (information issue de l'annotation manuelle par les rédacteurs). Depuis le stockage de ces informations dans une base de données, nous pouvons alors extraire n'importe quelle information sous n'importe quelle forme afin de mener divers tests statistiques.

La section suivante présente les stratégies mises en place pour mettre au jour les traits linguistiques susceptibles d'être de bons marqueurs de l'obsolescence.

6 Des statistiques pour justifier l'utilisation de données complexes

Parce que nous disposons d'une représentation des unités de discours sous un format matriciel, nous pouvons mener une série de tests statistiques afin de *(i)* comprendre et mesurer l'organisation des indices linguistiques dans les segments d'obsolescence annotés manuellement et *(ii)* mesurer l'impact des éléments linguistiques hiérarchiques et positionnels pour notre tâche.

Le choix de passer par les statistiques nous semble évident car il s'agit d'un outil approprié pour :

- traiter des données conséquentes (pour nous, matrice de 9916 individus et 150 variables) ;
- prouver/infirmer des hypothèses de recherche ;
- faire émerger des informations nouvelles.

Nous avons mené à la fois des statistiques de type descriptif et des statistiques prédictives. Dans les deux cas, cela permet de mieux comprendre et de décrire les segments d'obsolescence ainsi que les indices linguistiques fonctionnant en configuration et dans le second cas, cela permet de capitaliser des connaissances nouvelles pour les réutiliser sur de nouvelles données.

6.1 Statistiques descriptives

Afin de vérifier la pertinence des traits textuels susceptibles de jouer un rôle dans les segments d'obsolescence, nous avons mesuré dans un premier temps la corrélation de la variable *obsol* (interprétée à partir des annotations manuelles de l'obsolescence par les experts) avec chacune des autres variables de la base (les traits repérés de manière automatique par l'outil ALIDIS). Nous avons pour cela utilisé le logiciel SPAD [10].

Cette mesure de la pertinence des traits met au jour :

- des traits corrélés positivement [11] à la variable *obsol*, c'est-à-dire des traits qui apparaissent préférentiellement dans les segments d'obsolescence : par exemple, les entités

10. http://www.coheris.fr/fr/page/home.html

11. Le logiciel SPAD fournit une valeur test ou v-test. Les auteurs du logiciel indiquent que si la valeur-test est supérieure à 2, alors le cœfficient est significatif avec un risque d'erreur inférieur à 5% . Plus la valeur-test est grande

nommées de type *géopolitique* (« nombre d'habitants », « PIB » ; v-test à 26,00) ou de type *mesure évolutive* (« 30 hab./km2 » ; v-test à 24,39) ou encore les adverbiaux temporels de type *déictique coïncidence* (« aujourd'hui » ; v-test à 18,91) ,

– des traits corrélés négativement à la variable *obsol*, c'est-à-dire des traits qui apparaissent moins fréquemment dans les segments non-obsolescents : par exemple, les entités nommées de type *personne* (« Mohandas Karamchand Gandhi » ; v-test à -5,95) ou encore les adverbiaux temporels de type *ponctuel antériorité++* (« en 1800 » ; v-test à -6,18) ,
– des traits non corrélés à la variable *obsol*, c'est-à-dire des traits qui apparaissent indifféremment dans les segments obsolescents et dans les segments non-obsolescents : par exemple, le futur (v-test à 0,72).

Ces premiers calculs indiquent tout d'abord que les traits linguistiques pris en compte se justifient. En effet, sur les 150 variables étudiées, 53 sont corrélées positivement, 21 négativement et 76 sont neutres. Parce que nous recherchons à terme des combinaisons de traits textuels, nous faisons le choix de conserver pour les calculs qui vont suivre l'ensemble des traits, y compris ceux dont la corrélation avec la variable *obsol* est nulle. En effet, nous supposons qu'un trait, alors qu'il est insignifiant pour l'obsolescence lorsqu'il est seul dans un segment, est capable d'orienter l'interprétation obsolescente d'un segment s'il cooccurre avec un ou plusieurs autres traits.

Les corrélations mises à jour avec cette méthode confirment certaines de nos intuitions initiales. Par exemple, les adverbiaux temporels de type *déictique* (« aujourd'hui ») sont fréquemment corrélés à l'obsolescence.

Sur la base des résultats de corrélation, nous avons mis en place deux systèmes de base qui vont permettre la compararaison de différentes techniques d'apprentissage mais également de constituer une première méthode naïve de classification. Ces systèmes s'organisent ainsi :

Système 1 : partant du constat empirique que les traits les plus corrélés à l'obsolescence sont des valeurs numériques et des dates, nous testons tout d'abord si la simple présence de chiffres dans une phrase est suffisante pour déterminer sa nature obsolescente ou non ;

Système 2 : la seconde méthode de base est la présence d'au moins un des traits les plus corrélés à la variable *obsol* : 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/modes futur et conditionnel, les adverbiaux exprimant un point de vue de type récence (« les territoires actuels ») ou prévision (« les recherches à venir »).

Le tableau 3 renseigne sur les performances de ces systèmes de base, lesquelles sont relativement peu élévées.

	Précision	Rappel	F-score
Système 1	23	31	26
Système 2	30	39	37

TAB. 3 – *Performances des systèmes de base*

(en valeur absolue), plus la liaison entre variables est significative et moins le hasard a de chance d'être responsable de celle-ci.

Pour nous permettre d'automatiser le processus de détection des marqueurs complexes de l'obsolescence, nous avons testé, plusieurs systèmes d'apprentissage automatique supervisé.

6.2 Apprentissage supervisé

Un système d'apprentissage automatique permet d'extraire, sous forme de règles généralement, des régularités à partir d'une masse d'informations (Rioult et al., 2008). 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. Le fait que nous disposions d'un corpus annoté manuellement par des experts au niveau des segments d'obsolescence rend possible la mise en oeuvre d'une technique de classification supervisée : cela valide l'intérêt et la pertinence des indices dans les segments d'obsolescence et vérifie si une machine peut repérer automatiquement ces segments sur la base des indices présentés.

Méthode

Afin de déterminer quelle méthode est la plus adaptée à notre besoin, nous [12] avons effectué plusieurs expériences à l'aide de la plateforme RAPIDMINER [13]. Cette plateforme permet d'utiliser des méthodes classiques de classification supervisée :

- les arbres de décision (Quinlan, 1993; Cornuéjols et Miclet, 2002)
- un modèle probabiliste, le bayésien naïf (Hand et Yu, 2001)
- les séparateurs à vaste marge [14] (Boser et al., 1992)
- les réseaux de neurones (Minsky et Papert, 1969; Dreyfus et al., 2008)

Les configurations livrées en standard par RAPIDMINER fournissent immédiatement de bonnes performances. Ces résultats pourraient certainement être améliorés par un paramétrage adapté, mais requièrent une expertise que nous n'avons pas.

Le modèle à base de règles d'association que nous utilisons (Agrawal et al., 1993) extrait des expressions de la forme $X \rightarrow Y$ qui sont calculées à partir de motifs (*i.e.* des 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.

Le classifieur est réglé de la façon suivante :

- le nombre d'erreurs tolérées (le delta) est paramétré à 5, soit 5 erreurs tolérées par règle.
- le seuil de support est établi à 0.0001 (soit 1 pour 1000). Cela signifie que pour notre corpus, une règle doit fonctionner au moins sur 9 phrases (puisque le corpus contient 9916 phrases).

12. Merci à François Rioult pour son aide dans la mise en oeuvre de ces expériences (Laignelet et Rioult, 2010).
13. `http://www.rapidminer.com`
14. SVM - *support vector machine*

Résultats

Les résultats sont reportés au tableau 4.

Algorithme	AUC	Classe non obsolescent (85 %) précision	rappel	Fscore	Classe obsolescent (15 %) précision	rappel	Fscore
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
règles d'association	79.8	79.2	85.7	85.7	39.0	**70.3**	**50.1**

TAB. 4 – *Performances (en pourcentage) pour la classification supervisée (10-validation croisée).*

Les règles d'association semblent répondre le mieux à notre besoin essentiellement parce que nous voulons privilégier le rappel par rapport à la précision. En effet, pour notre problème, oublier une révision est plus grave qu'indiquer inutilement un segment à réviser. Il vaut donc mieux repérer trop de segments même s'ils ne sont pas obsolescents. Pour rappel, l'objectif est de proposer un outil d'aide à la mise à jour des contenus encyclopédique en annotant de manière automatique les segments susceptibles de contenir de l'information obsolète.

La vocation de ce type de technique est exploratoire : toutes les associations possibles sont générées qu'elle que soit la conclusion de classe produite. C'est dans un second temps qu'un tri est effectué sur la classe à étudier : nous focalisons notre attention sur les règles concluant sur la classe $obsol$. Ce type de technique permet également de valider les traits et la prégnance de la classe puisqu'elle émerge (ou non) naturellement des statistiques.

Exemple de règle d'association apprise

$$exprTemp.nature : deictique; sitTps : coincidence \wedge entiteNom.classe : mesure; sousClasse : evolutif \longrightarrow classe : obsol$$

Cette règle stipule que si un segment contient une expression temporelle de type *déictique - coïncidence* (« aujourd'hui ») et une entité nommée de type *mesure évolutive* (« 35 hab./km2 ») alors il est possible de conclure sur la classe $obsol$ et donc de supposer que le segment en question est obsolescent. L'exemple 6 illustre les types de phrases repérés par cette règle.

§ 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 6 – *Combinaison de plusieurs indices intra-phrastiques*

Les cas d'erreur (règles incorrectes)

Certaines phrases sont considérées comme obsolescentes suite au repérage erroné de certains indices par l'outil ALIDIS. Par exemple, dans l'exemple 7, c'est à cause d'une mauvaise 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 »). La granularité temporelle devrait être mieux prise en compte, plus fine et plus adaptée au problème de l'obsolescence : ici « notre ère » est d'une trop grande granularité par rapport aux besoins de mise à jour de l'information qui, généralement, sont nécessaires au maximum au siècle près.

Le gnomon : Le plus primitif des cadrans solaires date du IIe millénaire **avant notre ère**. [. . .]

Source : Corpus Atlas

Exemple 7 – *Erreur d'annotation de l'outil* ALIDIS

Les cas où une phrase contenant de l'information obsolescente n'est pas repérée concernent un cas sur trois. Comme pour les segments mal étiquetés, 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 8, 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 8 – *Pondérer fortement certains indices (temporels par exemple)*

Perspectives d'évolution

La difficulté majeure dans notre approche par apprentissage automatique concerne la taille du corpus et plus précisément le faible nombre de cas étiquetés comme obsolescents parmis l'ensemble des phrases.

Plusieurs solutions sont possibles (Gotab, 2009; Charton et al., 2008). Tout d'abord, l'active-learning (Riccardi et Hakkani-Tur) permetrait de contrebalancer la faible quantité de données d'apprentisssage tout en en augmentant la qualité. Une sélection des exemples permettrait parallèlement de mettre en place un corpus peu redondant et donc plus efficace informativement. Dans notre situation, augmenter la qualité des exemples d'apprentissage doit passer par une redéfinition de l'obsolescence et des besoins réels des utilisateurs finaux. Sur une partie du corpus (le sous corpus Larousse), nous disposons d'une multi-annotation : les cas d'accord absolu (*i.e.* pourraient soit être les seuls à être pris en compte, soit pondérés positivement par rapporot aux cas où le segment n'est annoté que par un expert.

Une seconde piste concerne le co-training (Blum et Mitchell, 1998). Cette méthode permet d'augmenter artificiellement un corpus d'apprentissage à travers l'ajout de données non annotées. Une telle approche mélant apprentissage supervisé et non supervisé est intéressant dans

notre cas, dans la mesure où nous disposons d'un corpus non annoté très important (l'ensemble de l'encyclopédie Larousse, soit 150 000 articles).

Malgré les limites évidentes de nos outils d'apprentissage et la nécessité de mener des expérimentations plus adaptées, les résultats montrent l'intérêt de considérer des données linguistiques complexes. Dans la section suivante, nous montrons que les performances de notre classifieur sont améliorées lorsque l'ensemble des traits textuels à granularité variable est pris en compte.

7 Les données complexes : quel impact ?

L'observation des règles d'association nous amène au constat suivant : tous les types de traits sont utilisés et toutes les associations sont représentées :

- des traits hiérarchiques avec des traits intra-phrastiques ;
- des traits positionnels textuels avec des traits intra-phrastiques ;
- des traits positionnels phrastiques avec des traits intra-phrastiques ;
- plusieurs traits intra-phrastiques ;
- enfin, l'importance du domaine pour certaines configurations.

En d'autres termes, toutes les granularités sont importantes et méritent d'être exploitées. Pour vérifier de manière plus précisément l'influence du grain d'analyse, nous avons mis en place une nouvelle expérimentation. Nous avons créé des vues différentes sur nos données :

- le *corpusComplet* est la vue décrite jusqu'ici, c'est-à-dire une vue qui prend en compte tous les traits ;
- le *corpusIPseuls* est une vue qui prend en compte uniquement les traits intra-phrastiques ;
- le *corpusIPHierar* est une vue qui prend en compte les traits intra-phrastiques et les traits hiérarchiques ;
- le *corpusIPPos* est une vue qui prend en compte les traits intra-phrastiques et les traits positionnels ;
- le *corpusEpure* est un corpus « épuré » dans lequel sont enlevées les variables non significatives (en fonction des résultats des statistiques descriptives). Tous les types de traits sont présents mais en nombre limité.

Comme cela a été fait précédemment sur le corpus entier, nous avons procédé à un apprentissage automatique sur chacun des vues présentées ci-dessus. Nous utilisons dans cette expérience un système à base de règles d'associations.

Dans le tableau 5 ci-dessous, les performances liées à chacune de ces vues sont indiquées.

Ces résultats montrent tout d'abord qu'exploiter les indices intra-phrastiques uniquement est insuffisant. Les mesures de précision et de rappel sont très basses (respectivement 38 % et 37 %).

En revanche, le gain est mesurable avec l'exploitation d'indices de type hiérarchique (+ 5 % sur le F-score), d'indices positionnel (+ 4.4 %) ou les deux (+ 8.9 %).

De plus, les indices hiérarchiques semblent favoriser la précision alors que les indices positionnels semblent privilégier le rappel.

Enfin, il semblerait qu'il vaille mieux ne pas considérer trop d'indices : les résultats associés au corpus épuré (qui contient tous les types d'indices qui ont été mesurés comme étant significatif dans les segments d'obsolescence) ont une meilleure précision que lorsque tous les indices sans distinction sont pris en compte.

	Précision	Rappel	F-Score
Base 1	23	31	26
Base 2	30	39	37
corpusComplet	32.9	**78.8**	46.4
corpusIPseuls	38	37	37.5
corpusIPHierar	**39.9**	45.6	42.5
corpusIPPos	33.2	56.7	41.9
corpusEpure	38.7	62.3	**47.7**

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

Par exemple, concernant le rôle des titres dans la recherche des segments obsolescents, nous observons que l'association entre des indices hiérarchiques et des indices intra-phrastiques est assez fréquente. Ainsi, comme l'illustre l'exemple 9 suivant, 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 9 – *Exemple de combinaisons d'indices hiérarchiques et d'indices intra-phrastiques*

8 Conclusion

Dans cet article, nous avons cherché à montrer dans quelle mesure il est possible et utile de dépasser l'unité phrastique lorsqu'on travaille sur le traitement automatique des textes. Nous avons présenté un certain nombre de structures linguistiques à la fois hétérogènes d'un point de vue sémantique et à granularité variable (du mot à la section en passant par les titres et les positions des éléments).

Parce que le discours n'est pas considérée comme une unité linguistique valide par une certaine communauté de linguistes (Reboul et Moeschler, 1996), il est essentiel, dans l'hypothèse de l'existence de telles unités discursives, de se donner les moyens de répondre aux principales objections, à savoir la définition des objets et de leurs relations.

Parallèlement à ce débat linguistique, un second enjeu, plutôt appliqué, apparait : les applications basées sur une analyse de la langue sont de plus en plus développées et requièrent de plus en plus de connaissances "profondes" sur la langue. C'est ainsi que, pour résoudre notre tâche de repérage de l'obsolescence, des indices linguistiques "simples" ne suffisent pas.

Face à ce double objectif, nous avons proposé un modèle de représentation des unités discursives. Il s'agit de se donner les moyens de transformer des informations linguistiques hétérogènes en un format analysable par des outils statistiques au sens large. Bien entendu, notre proposition est imparfaite et ne modélise que partiellement la réalité des discours. En effet, nous n'envisageons les relations entre unités qu'en termes de hiérarchie, alors que dans les discours les relations horizontales (entre phrases par exemple comme avec l'utilisation de pronoms) sont essentielles. Notre modélisation reste une proposition qui doit être améliorée notamment pour rendre compte des relations pronominales, des chaines référentielles ou encore des associations sémantiques.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *Proc. of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, USA*, pp. 207–216.

Biber, D., U. Connor, et T. Albin-Upton (2007). *Discourse on the move : using corpus analysis to describe discourse structure*. John Benjamins.

Blum, A. et T. Mitchell (1998). Combining labeled and unlabeled data with co-training. In *COLT' 98 : Proceedings of the eleventh annual conference on Computational learning theory*, New York, NY, USA, pp. 92–100. ACM.

Boser, B. E., I. M. Guyon, et V. N. Vapnik (1992). A training algorithm for optimal margin classifiers. In *5th Annual ACM Workshop on COLT*, pp. 144–152.

Bouillon, P., C. Fabre, P. Sébillot, et L. Jacqmin (2000). Apprentissage de ressources lexicales pour l'extension de requêtes. *TAL : Traitement Automatique des Langues 41*(2), 367–393.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Morristown, NJ, pp. 152–155. Association for Computational Linguistics.

Brunet, E. (2003). Peut-on mesurer la distance entre deux textes ? *Corpus* (2).

Busquets, J., L. Vieu, et N. Asher (2001). La sdrt : une approche de la cohérnce du discours dans la tradition de la sémantique dynamique. *Verbum XXIII*(1), 73–101.

Carter-Thomas, S. (2001). *La cohérence textuelle*. L'harmattan.

Charolles, M. (1997). L'encadrement du discours, univers, champs, domaine et espaces. *Cahiers de Recherche linguistique 6*.

Charton, E., N. Camelin, R. Acuna-Agost, P. Gotab, R. Lavalley, R. Kessler, et S. Fernandez (2008). Pré-traitements classiques ou par analyse distributionnelle : application aux méthode de classification automatique déployées pour deft08. In *Actes de TALN 2008*.

Cornuéjols, A. et L. Miclet (2002). *Apprentissage artificiel : Concepts et algorithmes*. Eyrolles.

Daille, B., C. Fabre, et P. Sébillot (2001). Applications of computational morphology. In P. Boucher (Ed.), *Many morphologies*, Somerville, Massachussetts, pp. 210–234. Cascadilla Press.

Degand, L. et T. Sanders (2002). The impact of relational markers on expository text comprehension in l1 and l2. *Reading and Writing 7-8*(15), 739–758.

Dejean, C., M. Fortun, C. Massot, V. Pottier, F.Poulard, et M. Vernier (2010). Un étiqueteur de rôles grammaticaux libre pour le francais intégré à apache uima. *Actes de TALN 2010*.

Dreyfus, G., J.-M. Martinez, M. Samuelides, M. Gordon, F. Badran, et S. Thiria (2008). *Apprentissage statistique : réseaux de neurones, cartes topologiques, machines à vecteurs supports*. Eyrolles.

Enjalbert, P. (2005). *Sémantique et TALN*. Hermes.

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

Gotab, P. (2009). Apprentissage automatique et co-training. In *Actes de la conférence RECITAL 2009*.

Grabar, N. et P. Zweigenbaum (1999). Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In *Actes de la VIe conférence sur le Traitement Automatique des Langues Naturelles (TALN '99)*, Paris, pp. 175–184. Talana, Université Paris 7.

Halliday, M. et R. Hasan (1976). *Cohesion in English*. London : Longman Group Limited.

Hand, D. et K. Yu (2001). Idiot's bayes - not so stupid after all ? *International Statistical Review 69*(3), 385–399.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1992)*, Morristown, NJ, pp. 539–545. Association for Computational Linguistics.

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

Hripcsak, G. et D. Heitjan (2002). Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics 35*(2), 99–110.

Jackiewicz, A. (2005). Les séries linéaires dans le discours. *Langue Française 148*, pp. 95–110.

Jacquemin, C. (1997). Guessing morphology from terms and corpora. In P. Willett et A. D. Narasimhalu (Eds.), *Proceedings of the 20th Annual International ACM SIGIR Conference*

on Research and Development in Information Retrieval (SIGIR'97), New York, pp. 156–165. Association for Computing Machinery.

Johnson, K. et H. Johnson (1999). *Encyclopaedic dictionary of applied linguistics*, Chapter XX, pp. 55–57, 99–101. Oxford : Blackwell Publishers Ltd.

Kao, A. et S. Poteet (Eds.) (2007). *Natural Language Processing and Text Mining*. Springer.

Laignelet, M. (2009). *Analyse discursive pour le repérage automatique de segments obsolescents dans les documents encyclopédiques*. Ph. D. thesis, Université de Toulouse - Le Mirail.

Laignelet, M. et F. Rioult (2010). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. *TAL 51*(1), 41–63.

Lavallée, J.-F. et P. Langlais (2010). Apprentissage non supervisé de la morphologie d'une langue par généralisation de relations analogiques. *Actes de TALN 2010*.

Li, W., J. Han, et J. Pei (2001). Cmar : Accurate and efficient classification based on multiple class-association rules. In *IEEE International Conference on Data Mining*.

Mani, I. (2001). *Automatic summarization*. Amsterdam/Philadelphie : John Benjamins Publishing Compagny.

Mann, W. C. et S. Thompson (1988). Rhetorical structure theory : toward a functional theory of text organisation. *Text 8*(3), 243–281.

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

Nouvel, D., A. Soulet, J.-Y. Antoine, N. Friburger, et D. Maurel (2010). Reconnaissance d'entités nommées : enrichissement d'un système à base de connaissances à partir de techniques de fouille de textes. *Actes de TALN 2010*.

Péry-Woodley, M.-P. (2000). Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle. Technical report, Carnet de grammaire, Rapport n8. thèse d'habilitation.

Péry-Woodley, M.-P. (2005). *Sémantique et Corpus*, Chapter Discours, corpus, traitements automatiques, pp. 177–210. Paris : Hermes Science - Lavoisier.

Péry-Woodley, M.-P., N. Asher, P. Enjalbert, F. Benamara, M. Bras, C. Fabre, S. Ferrari, L.-M. Ho-Dac, A. LeDraoulec, Y. Mathet, P. Muller, L. Prévot, J. Rebeyrolle, L. Tanguy, M. V.-C. M., L. V. L., et A. Widlöcher (2009). Annodis : une approche outillée de l'annotation de structures discursives. *TALN 2009*.

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

Reboul, A. et J. Moeschler (1996). Faut-il continuer à faire de l'analyse de discours ? *Journal of Linguistics 16*.

Reichenbach (1966 (1ère édition 1947)). *Elements of symbolic logic New- york*. Free-Press.

Riccardi, G. et D. Hakkani-Tur. Active learning : Theory and applications to automatic speech recognition.

Rioult, F., B. Zanuttini, et B. Crémilleux (2008). Apport de la négation pour la classification supervisée à l'aide d'associations. In *Conférence d'Apprentissage*, pp. 183–196.

Rioult, F., B. Zanuttini, et B. Crémilleux (2010). *Advances in Intelligent Information Systems*, Volume 265 of *Studies in Computational Intelligence*, Chapter Nonredundant generalized rules and their impact in classification, pp. 3–25. Springer.

Sarfati, G.-E. (1997). *Eléments d'analyse du discours*. Paris : Nathan.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In D. Jones (Ed.), *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, pp. 44–49. UMIST.

Sébillot, P., P. Bouillon, V. Claveau, C. Fabre, L. Jacquemin, et J. Nicolas (2000). Apprentissage en corpus de couples nom-verbe pour la construction d'un lexique génératif. In *Actes des JADT 2000 (Journées d'Analyse de Données Textuelles)*, Lausanne, pp. 205–212. Ecole Polytechnique Fédérale de Lausanne.

Teufel, S. (1998). Meta-discourse markers and problem-structuring in scientific articles. In *Workshop on Discourse Structure and Discourse Markers*, Montreal. ACL 1998.

Teufel, S., J. Carletta, et M. Moens (1999). An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL*.

Widlöcher, A. et F. Bilhaut (2005). 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.

Zerida, N., N. Lucas, et B. Crémilleux (2006). Combinaison de descripteurs linguistiques et de structure pour la fouille d'articles biomédicaux. In *Schedae*, pp. 69–78.

Summary

This paper presents a methodology for the processing and the building of complex linguistic data. A complex data corresponds to textual information that is semantically heterogeneous and whose scope is variable. We propose a model in order to represent the discourse units : such a model permits to transform a linguistic structure constituted by complex objects in a representation of datas that is used for statistical tools. This work is led into a research project which aim is to develop a prototype for helping human in update tasks of encyclopedic texts. More precisely, the goal is the automatic identification of textual segments which contain information requiring updating ("obsolescence segments").