Repérage de segments d'information évolutive dans des documents de type encyclopédique

Marion Laignelet[1,2]

[1]Université de Toulouse le Mirail, ERSS, UMR 5610
[2]Société Initiales
marion.laignelet@univ-tlse2.fr

Résumé

Dans cet article, nous cherchons à caractériser linguistiquement des segments textuels définis pragmatiquement, relativement à des besoins de réédition de documents et au sein desquels l'information est susceptible d'évoluer dans le temps. Sur la base d'un corpus de textes encyclopédiques en français, nous analysons la distribution de marqueurs textuels et discursifs et leur pertinence en nous focalisant principalement sur un traitement sémantique particulier de la temporalité.

Mots-clés : repérage automatique d'informations évolutives, TAL, corpus de textes encyclopédiques, marqueurs textuels et discursifs.

Abstract

In this paper, we present a method for the automatic identification of text segments carrying information likely to evolve in the course of time. In a corpus of extracts from French encyclopediae, we analyse relevant text and discourse markers focussing on a specific approach to temporal semantics.

Keywords: automatic identification of evolving information, NLP, corpus of encyclopaedic texts, textual and discourse markers.

1. Introduction

Cette étude s'inscrit dans un projet de recherche et développement[1] dont l'objectif applicatif est la création d'un prototype logiciel d'aide à la mise à jour de contenus encyclopédiques pour l'édition. Plus spécifiquement, l'outil final devra repérer automatiquement des zones de textes susceptibles de contenir de l'information obsolescente, qui évolue dans le temps. Ces zones, ou segments d'information évolutive, se définissent donc d'abord par rapport à un usage spécifique (la tâche de mise à jour éditioriale). Nous cherchons à définir linguistiquement ces éléments définis en fonction d'un usage. Pour ce faire, nous exploitons un certain nombre de méthodes et techniques issues de la linguistique de texte et de discours, de la linguistique de corpus ou encore du TAL (extraction d'information ou du résumé automatique). Le traitement de la composante temporelle des textes est au centre de nos recherches. Cependant, nous ne nous situons ni dans la lignée des travaux sur le calcul de la référence temporelle ou celui des intervalles temporels, ni dans le cadre d'une sémantique de contenu. Ce travail exploite certes la composante sémantique des manifestations temporelles présentes à la surface des textes mais en vue d'un

[1] Collaboration entre la société d'édition Initiales et l'ERSS.

traitement applicatif spécifique et donc pour un usage particulier. Nous exploitons également la notion de marqueurs textuels et discursifs comme les « mots-repères »ou les « mots-titres », notions déjà envisagées par (Edmundson, 1969), les *cue phrases* (Grosz et Sidner, 1986) ou encore les éléments participant de l'analyse de la structure de texte (Marcu, 2000). De plus, considérant le caractère multifonctionnel des marqueurs de surface (Grosz et Sidner, 1986), nous nous focalisons sur leur fonction pragmatique, *i.e*. leur aptitude à déterminer un segment d'information évolutive. Enfin, les aspects discursifs des documents à travers les titres (Ho-Dac *et al*., 2004) ou les cadres de discours (Charolles, 1997) occupent également une place centrale dans nos recherches. Ce niveau discursif est depuis peu pris en compte dans les sytèmes de TAL (Crispino *et al*., 1999 ; Berri *et al*., 1996).

Dans un premier temps, nous définissons ce qu'est (pragmatiquement) un segment d'information évolutive sur la base d'un exemple extrait de notre corpus. Puis, s'agissant d'une étude en corpus, nous présenterons le matériel textuel utilisé servant pour l'analyse de la tâche, comme corpus d'étude et comme corpus d'évaluation. La méthodologie (les marqueurs de surface pris en considération et leur implémentation) adoptée sera ensuite décrite puis les résultats de cette étude et les perspectives qu'ils entraînent.

2. Qu'est ce qu'un segment d'information évolutive ?

Un segment d'information évolutive ou SEDIS-ε[2] est un segment textuel susceptible de contenir une ou plusieurs informations qui présente(nt) cette particularité de pouvoir évoluer dans le temps et/ou qui relativement à des besoins éditoriaux nécessiterai(en)t d'être réactualisé(es). L'exemple suivant illustre ce que nous entendons par « SEDIS-ε ».

(1) **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 de Nairobi (Kenya) ont ouvert de nouvelles perspectives. [...]. La recherche se tourne justement aujourd'hui vers des vaccins qui déclencheraient une réponse du système immunitaire à ces deux niveaux (anticorps et cellules tueuses). [...]. Des expériences ont été faites pour déterminer la nature de l'éventuel vaccin [...] mais qui n'est pas actuellement envisageable chez l'homme, [...]. En juin 2003, une équipe de biologistes américains a obtenu des résultats qui pourraient laisser envisager, à terme, l'élaboration d'un vaccin efficace. Les chercheurs sont parvenus, [...]. Cette découverte pourrait aboutir à la mise au point d'un antigène ([...]) qui aurait pour conséquence la création de l'anticorps 2G12, [...].

Dans cet exemple, l'auteur exprime une issue possible et probable concernant les recherches sur le sida. Il convient de préciser que la fiche de laquelle est extrait ce passage a été éditée et distribuée dans le courant de l'année 2003 et que ce type de fiche est destiné à être distribué dans le cadre d'éditions dites « au long cours »[3] : un client peut s'abonner à l'encyclopédie en 2007 et est susceptible de recevoir la fiche écrite en 2003 et dans laquelle cet exemple apparaît ; cependant, pendant ce laps de temps (de 2003 à 2007), soit certaines des prédictions formulées par l'auteur se seront réalisées, soit elles auront été repoussées par les scientifiques

[2] De SEgment de DIScours et ε pour référer à la notion d'information évolutive.

[3] Ce sont des éditions fonctionnant sous forme d'abonnement ; le client s'abonne à un moment T et pendant une durée déterminée, il va recevoir un nombre déterminé de fiches tous les mois ; ce type d'édition dure en général entre 5 et 7 ans voire plus si la collection fonctionne bien.

ou encore de nouvelles données peuvent entrer en jeu. Il est donc tout à fait souhaitable que ce segment de texte ait été préalablement mis à jour. À travers cet exemple, nous pouvons observer l'importance de la composante temporelle. L'objectif de cette étude n'est ni de procéder à des calculs de la référence temporelle comme (Aurnague *et al*., 2001) par exemple, ni de chercher à associer un événement particulier à une date particulière comme c'est le cas en extraction d'information ou dans les systèmes de question-réponse. Nous ne cherchons, par exemple, ni à valider le fait que l'événement *« une équipe de biologistes américains a obtenu des résultats qui [...] »* est vrai, ni qu'il s'est réellement produit en *« juin 2003 »*. Nous recherchons les segments pour lesquels il est pertinent de penser que l'information donnée est susceptible d'avoir évolué entre le moment de l'édition de la fiche et le moment de réédition. Dans l'exemple, l'information *« Cette découverte pourrait aboutir à la mise au point d'un antigène »* correspond au type de segment recherché. Nous pouvons dès à présent mentionner l'existence de deux types de SEDIS-ε. Comme nous venons de le voir à travers l'exemple 1, il existe des segments textuels qui nécessitent une ***mise à jour*** : l'information n'est plus vraie ou ne s'est pas vérifiée (c'est le cas lorsque l'auteur fait des prédictions dans le futur sur un fait ou un événement). Il existe également des segments contenant une information, qui, dans l'absolu, restera toujours vraie (par exemple le fait que *le PIB de la France s'élève à 1 320,4 milliards de dollars en 2001*) mais qui, dans un contexte éditorial, nécessite de faire l'objet d'une ***réactualisation*** (*les chiffres du PIB de l'année en cours*).

3. Le corpus

Le corpus d'étude est constitué, à ce jour, de 92 textes de type encyclopédique[4]. Il s'agit de fiches encyclopédiques éditées et accessibles sur le marché de l'édition (propriété des Editions Atlas). A priori donc, du point de vue du type de texte (cf. terminologie de (Biber, 1988 ; Biber, 1989)) le corpus est homogène. Le trait distinctif entre ces textes relève de la catégorisation en genre et plus précisément du domaine de connaissance auquel les fiches appartiennent. Nous insistons sur ce point parce que nous supposons l'importance de cette distinction par domaine pour les résultats. Huit domaines différents sont représentés : géographie (14), médecine & santé (13), sciences & techniques (10), société (8), sport (8), histoire (17), art & littérature (12) et faune & flore (7). Nous envisageons à terme d'augmenter la taille de ce corpus. Ce corpus [ATLAS] constitue une base de 80 000 mots, au format xml.

4. Méthodologie

Une première lecture de l'ensemble du corpus nous a permis d'inventorier (de manière non exhaustive) un certain nombre de marqueurs textuels et discursifs qui nous sont apparus pertinents et aptes à délimiter des SEDIS-ε. Parallèlement, nous avons annoté manuellement, pour une partie du corpus seulement, les zones textuelles étant des SEDIS-ε (cf. chapitre 4.1). Nous avons ensuite projeté, de manière automatique (cf. chapitre 4.3), l'ensemble des marqueurs considérés, ainsi que des informations sémantiques liées (cf. chapitre 4.2), sur le corpus annoté manuellement afin d'évaluer quantitativement ceux qui effectivement sont présents dans un SEDIS-ε et donc pertinents pour la tâche. Si nous procédons de cette manière, c'est que nous sommes encore dans une phase exploratoire au cours de laquelle nous cherchons à exploiter et à rendre compte de la fonction pragmatique de certains marqueurs textuels et discursifs, et ce, malgré leur caractère multifonctionnel dans les textes.

[4] Nous ferons référence à ce corpus par [ATLAS].

4.1. Annotation manuelle d'une partie du corpus

Pour environ la moitié du corpus [ATLAS][5], nous avons procédé à une annotation manuelle des SEDIS-ε. Cette tâche a consisté à marquer (manuellement et par balisage xml) les segments de textes [6] dans lesquels un annotateur (non expérimenté) jugeait que l'information contenue était susceptible d'être mise à jour. Le fait que l'annotateur soit inexpérimenté pour la tâche demandée a entraîné un certain nombre d'incohérences dans l'annotation ; c'est pourquoi nous envisageons de mettre en place un système d'annotation « multi-annotateur »tel que défini dans (Ferrari *et al*., 2005) ainsi qu'un protocole d'annotation manuelle précis.

4.2. Les marqueurs de surface considérés

Le choix des marqueurs de surface pris en compte fait suite à une observation et un relevé manuels minutieux du corpus entier. Ce relevé préalable a placé au centre de notre recherche les aspects temporels. D'autres éléments plus diversifiés sont également pris en considération : il s'agit de certaines expressions habituellement classées sous le terme d'entité nommée comme les sigles (en Extraction d'Information notamment), ou encore d'éléments participant de la structure argumentative des textes. Enfin, sont également exploités les aspects discursifs des documents à travers notamment la présence des marqueurs temporels à l'initiale de la phrase ou dans les titres. Le tableau 1 donne un aperçu des indices textuels et discursifs pris en considération et susceptibles de fonctionner comme marqueurs de SEDIS-ε.

4.3. Implémentation

Le repérage des marqueurs de surface est effectué de manière automatique à l'aide de la plate-forme LinguaStream[7] (Widlöcher et Bilhaut, 2005). LinguaStream est une plate-forme générique pour le traitement automatique des langues qui permet d'effectuer des traitements et des analyses de types et de niveaux linguistiques variés (morphologique, syntaxique, sémantique, discursif ou encore statistique) sur des corpus en XML. Dans le tableau 1, nous avons indiqué quel type de formalisme est utilisé pour repérer et annoter automatiquement chacun des indices susceptibles de marquer un SEDIS-ε. Basé sur le langage XML, LinguaStream nous permet également de travailler directement sur notre corpus annoté manuellement des SEDIS-ε. Les divers indices, qui sont potentiellement des marqueurs de SEDIS-ε, sont ainsi projetés sur le corpus annoté manuellement. Il nous est alors possible d'observer quels indices apparaissent effectivement dans un SEDIS-ε et de quantifier leur distribution (à l'intérieur d'un SEDIS-ε ou non).

Les diverses possibilités de visualisation des résultats constituent également un avantage certain (cf. figure 2). Ainsi, à l'issue de la chaîne de traitement de LinguaStream, nous disposons d'un

[5] Principalement pour des raisons de temps, nous n'avons, à ce jour, annoté manuellement qu'une partie du corpus. Par domaines, le nombre de fiches annotées manuellement est : géographie (10/14), médecine & santé (11/13), sciences & techniques (8/10), société (6/8), sport (3/8), histoire (0/17), art & littérature (0/12) et faune & flore (0/7). Cette tâche d'annotation va être effectuée sur la totalité du corpus. Il est important d'ajouter que les expérimentations présentées dans cet article ont été menées uniquement sur les fiches annotées manuellement.

[6] Le grain minimal considéré entre deux balises est la phrase sauf pour quelques cas particuliers : *« Dans les premières années de la recherche, le risque de contamination était d'environ 25 %, <DEBUT de* SEDIS-ε*> il est aujourd'hui de moins de 2 % <FIN de* SEDIS-ε*> »*

[7] http ://www.linguastream.org

Types d'indice	Traits sémantiques et *Valeurs*	Implémentation
Syntagme nominal temporel	année : *expression du texte* ; grain : *expression du texte*	Lexiques + Grammaire EDCG
Adverbial temporel	Eléments sémantiques de 'sntemps', borne : *totalité, début, fin,...* ; situation temporelle : *antériorité, postériorité, coincidence,...* ; type : *déictique ou non*	Lexiques + Grammaire EDCG
Temps verbaux	temps : *présent, passé-simple, passé composé, futur simple,...*	Tree tagger + Grammaire EDCG
Périphrases	type de modalité : *accomplissement, déroulement, répétition,...*	Macro expressions régulières
Argumentation	type : *correction, explication, opposition, conséquence, temporelle, exemplification,...*	Lexique
Sigle	*expression du texte*	expressions régulières
Superlatifs	type : *le plus* / *le moins* ; quoi : *expression du texte*	Macro expressions régulières
Valeurs chiffrées	type : *km, hab.,...*	Expressions régulières
Position Introducteurs de cadre	*OUI* / *NON*	Macro expressions régulières
Position dans un titre	*OUI* / *NON*	Macro expressions régulières

Tableau 1. Quelques indices textuels et discursifs

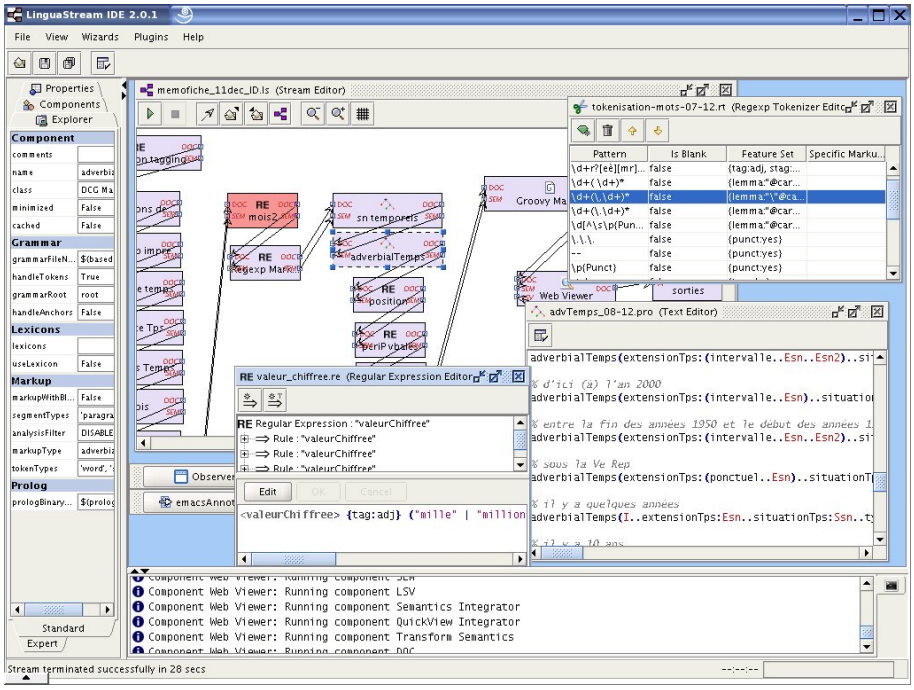

Figure 1. Chaîne de traitement avec la plateforme LinguaStream

document XML dans lequel nous pouvons exploiter et analyser les indices et leur représentation symbolique (sous la forme d'une structure de traits) décrits dans le tableau 1 à l'intérieur des SEDIS-ε annotés manuellement.

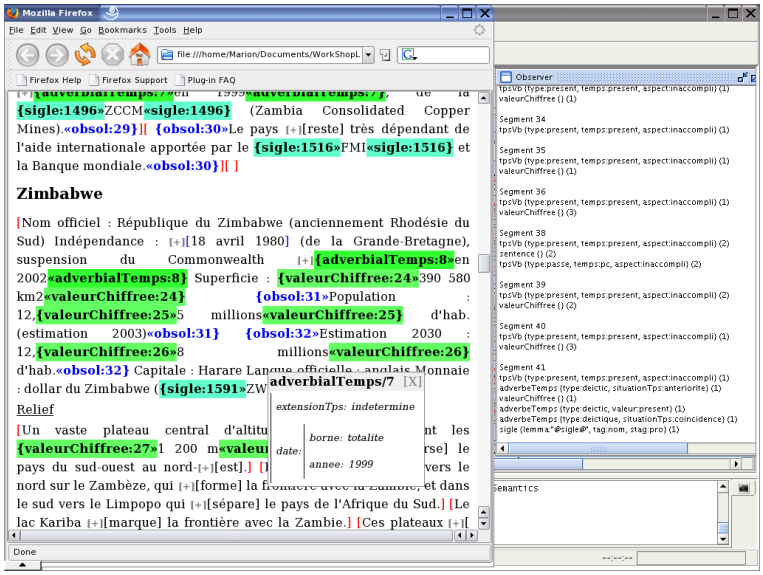

Figure 2. Exemples de sorties avec LinguaStream

5. Résultats : les marqueurs pris individuellement

Comme nous l'avons déjà mentionné, notre démarche a consisté à projeter l'ensemble des marqueurs potentiellement aptes à apparaître dans un SEDIS-ε sur le corpus dans lequel nous avons annoté manuellement les SEDIS-ε. Nous avons ensuite observé et analysé la distribution des marqueurs apparaissant ou non dans un SEDIS-ε puis nous avons quantifié leurs occurrences. Le tableau 2 permet d'observer le nombre et la proportion effective de SEDIS-ε au sein du corpus[8]. Nous présentons les résultats par domaine car des différences intéressantes émergent. La première remarque concerne le sous-corpus GÉOGRAPHIE pour lequel plus de 10 % du document est potentiellement à mettre à jour contre 3 % environ pour le domaine SOCIÉTÉ. Une seconde remarque importante concerne le fait que pour 16,40 % des SEDIS-ε annotés manuellement, aucun indice pris en compte n'émerge pour les caractériser en tant que tels. Deux raisons, principalement liées au caractère exploratoire de cette étude, expliquent ce résultat : premièrement, nous avons relevé un certain nombre d'erreurs dans l'annotation manuelle des SEDIS-ε ; deuxièmement, le programme de tokenisation ainsi que celui ayant pour but de relever les valeurs chiffrées présentaient, au moment de la rédaction de l'article, un certain nombre d'erreurs qui ont eu une influence certaine sur les résultats. Enfin, il est évident que la liste des indices textuels et discursifs est actuellement trop succincte et devra être enrichie quantitativement et qualitativement.

Le tableau 3 présente la répartition de quelques indices à l'intérieur des SEDIS-ε ou non. Conformément à nos intuitions, les aspects temporels occupent une place déterminante, dans deux cas : lorsque l'expression adverbiale est déictique, ou bien lorsqu'elle fait référence à une date ponctuelle proche du moment de lecture (en l'occurrence l'année 2005)[9]. Il en est de même pour les expressions temporelles marquant un intervalle de temps borné à son initiale (*« Depuis 1997 »*). Concernant les expressions temporelles (adverbe, adverbial ou syntagme nominal) apparaissant dans un titre ou à l'initiale de la proposition (et donc susceptibles d'avoir une fonc-

[8] Les *** signifient que les données sont faussées : en l'occurrence deux de ces fiches présentant des résultats sportifs les uns à la suite des autres, la notion de phrase ne revêt pas la même signification que pour les autres domaines (énumération de chiffres, de noms de personnes, pas de verbes, etc).

[9] Nous avons considéré un intervalle allant de 1997 à 2005.

	Nombre de SEDIS-ε dans le sous-corpus annoté manuellement	Proportion de phrases incluses dans un SEDIS-ε	Proportion de SEDIS-ε "vides"
GÉOGRAPHIE	376	**10,11 %**	**22,87 %**
MEDECINE & SANTÉ	154	5,36 %	**3,90 %**
SCIENCES & TECHNIQUES	138	5,10 %	12,32 %
SOCIÉTÉ	111	**3,36 %**	16,25 %
SPORT	20	***	20 %
TOTAL	799	6,30 %	16,40 %

Tableau 2. Proportion de SEDIS-ε *dans le sous-corpus annoté manuellement*

	Occurrences à l'intérieur d'un SEDIS-ε	Occurrences dans l'ensemble du corpus	Pourcentage
AdverbeTemps/ Déictique	126	238	**52,94 %**
AdverbialTemps/ année >1997	83	237	35,02 %
AdverbialTemps/ déictique	11	16	**68,75 %**
AdverbialTemps/ borneDébut	43	98	43,88 %
Futur	42	143	29,37 %
Conditionnel	83	236	35,17 %
Présent	672	6440	**10,42 %**
Périphrases	5	9	**55,55 %**
Superlatif	48	178	26,97 %
Temps/InTitre	2	7	28,57 %
Temps/IC	19	135	14,07 %

Tableau 3. Marqueurs présents dans les SEDIS-ε

tion cadrative), les chiffres montrent que les aspects cadratifs ont un rôle secondaire dans la détermination de SEDIS-ε. À ce sujet, une étude spécifique sur la nécessité de prendre en considération les aspects discursifs pour le repérage des SEDIS-ε est actuellement en cours, (cf. l'exemple de la figure 4). Les résultats concernant les périphrases verbales considérées sont encourageants, d'autant plus que, dans la plupart des cas, il s'agit de périphrases référant à une action dont l'accomplissement est en cours. Les temps verbaux n'apparaissent pas, à travers ces chiffres, comme de « bons »marqueurs de SEDIS-ε : le futur et le conditionnel, deux temps pour lesquels nous supposions une implication forte, ont des pourcentages d'apparition dans les SEDIS-ε relativement faibles. Ces résultats sur les temps verbaux nous amènent à diriger nos futures recherches vers un traitement des changements de temps verbaux plutôt que sur leur analyse locale et isolée. Enfin, les superlatifs ont un rôle important (presque 27 %) ; cependant, il sera à terme nécessaire de les distinguer plus finement. Par exemple, un superlatif comme *« le plus haut sommet du monde »* a peu de chances d'impliquer une évolution de l'information qui va suivre, alors qu'un superlatif du type *« le pays le plus peuplé »* est susceptible d'indiquer une évolution de l'information liée. Une dernière remarque concerne le cas des valeurs chiffrées. Nous considérons ces éléments à la fois comme des indices textuels pour le repérage de SEDIS-ε et en même temps comme des zones (très locales) susceptibles de devoir être mises à jour. Ce type d'information est, parmi l'ensemble des indices considérés, celui qui est le plus susceptible d'évolution dans le temps : effectivement, les résultats montrent l'importance de ce type de marqueur, et ce pour trois des domaines étudiés : en GÉOGRAPHIE (69,80 %), en MEDECINE & SANTÉ (76,07 %) et en SOCIÉTÉ (80 %). Nous envisageons un traitement plus fin et une anal-

yse sémantique plus poussée en distinguant des valeurs chiffrées du type "*nombre d'habitant*" ou "*PIB*", qui appellent des chiffres évolutifs, de celles pour lesquelles un changement est très peu probable (par exemple la superficie des pays).

6. Conclusion et perspectives

Pris individuellement, les marqueurs textuels et discursifs présentés sont insuffisants pour délimiter des SEDIS-ε. Même si les résultats décrits sont encourageants, il semble incontournable de les envisager en termes de configurations de traits plutôt que de manière isolée. Sur la base de deux exemples concrets extraits de notre corpus, voyons l'intérêt que représente une étude en configuration.

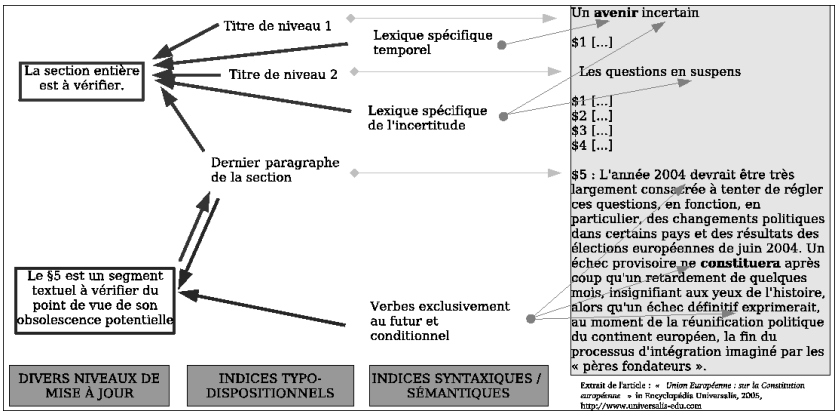

Figure 3. Exemple de SEDIS-ε *nécessitant une mise à jour*

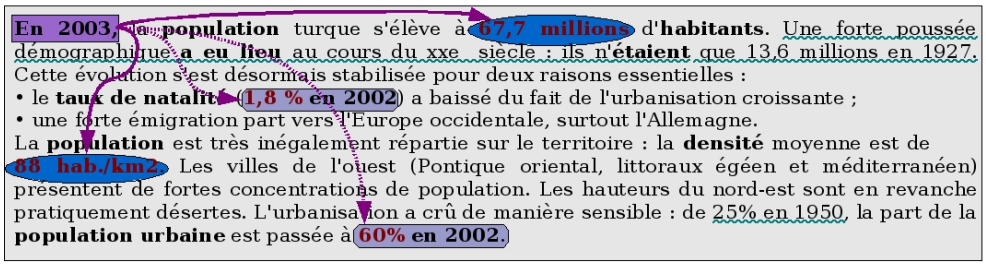

Figure 4. Exemple de SEDIS-ε *nécessitant une réactualisation*

Dans l'exemple de la figure 3, l'auteur fait mention d'une hypothétique issue concernant *l'année 2004* : il fait des prédictions sur une période qui, au moment où il écrit, n'a pas encore eu lieu. Plusieurs marqueurs de surface fonctionnent ici comme des signaux indiquant au lecteur que ce qui est écrit est virtuel, hypothétique : il s'agit notamment des temps verbaux (futur et conditionnel), du lexique présent dans les titres (*« avenir »*, ...) mais aussi du syntagme nominal *« L'année 2004 »* qui, si on se réfère à la date d'aujourd'hui (*2005*) est une période passée. Ce segment à mettre à jour est d'une longueur relative : le paragraphe voire la section entière si on traite les aspects discursifs tels que les titres. De plus, aucun élément en particulier ne semble devoir être mis à jour, c'est le segment dans son ensemble qui doit subir des modifications (cf. chapitre 2). La différence de l'exemple 4 (figure 4) par rapport à l'exemple 3 (figure 3) est d'abord liée au fait que les mises à jour à produire sont ici très locales (c'est ce que nous avons

nommé « réactualisation »dans le chapitre 2) : la forme générale du paragraphe ne nécessite pas de modification, mais seulement les dates et les valeurs chiffrées données. En effet, l'association entre une référence temporelle (par exemple *« en 2002 »*) et une valeur chiffrée (par exemple, *« 1,8 % »* et *« taux de natalité »*) restera toujours vraie (tout comme le fait qu'*« en 1950 l'urbanisation était de 25 % »*) ; cependant, il est nécessaire, en vue d'une réédition, de réactualiser les informations données. Toutefois, soulignons la difficulté à distinguer des cas comme *« de 25 % en 1950, la part de la population urbaine est passée à 60 % en 2002 »* où la première valeur (*celle de 1950*) ne doit pas être modifiée alors que la seconde (*celle de 2002*) pourrait nécessiter de l'être. À noter également l'importance de l'introducteur de cadre *« en 2003 »* qui étend sa portée sur l'ensemble du paragraphe et qui est nécessaire pour l'interprétation des valeurs mises en valeur dans les ovales.

Traiter conjointement plusieurs marqueurs textuels et discursif apparaît, à travers les deux exemples que nous venons de décrire, comme une étape nécessaire. Non seulement la co-présence de certains marqueurs devra être analysée de manière précise, mais également le fait que certains indices sont de meilleurs marqueurs de SEDIS-ε que d'autres. Par exemple, un adverbial comme *« actuellement »*, même seul, sera un bon marqueur de l'évolution et l'ensemble de la phrase dans laquelle il apparaît pourra être considéré comme un SEDIS-ε ; par contre, un futur seul ne marquera pas nécessairement un SEDIS-ε, il devra être entouré d'autres marqueurs. Dans les cas de « mises à jour » (cf. exemple dans la figure 3), la présence de plusieurs marqueurs peut créer un phénomène de « contamination » (ascendante et/ou descendante) et impliquer ainsi de considérer de larges zones textuelles à mettre à jour (des sortes de « macro-SEDIS-ε ») composées de plusieurs SEDIS-ε.

Nous envisageons d'évaluer la pertinence des marqueurs en les projetant sur de nouveaux corpus de textes encyclopédiques non annotés préalablement. Nous pourrons ainsi vérifier la validité et la pertinence de nos marqueurs et évaluer l'intérêt potentiel des configurations de marqueurs. Nous devrons également vérifier si les marqueurs que nous considérons détectent aussi des zones textuelles qu'il ne faudrait pas mettre à jour. Cependant, il ne faut pas oublier le but applicatif de notre projet, et même si nos programmes détectent trop de segments, le gain temporel apporté pour la tâche visée doit rester au premier plan.

Remerciements

Antoine WIDLÖCHER, Josette REBEYROLLE, Marie-Paule PERY-WOODLEY, Marie-Paule JACQUES, Mai HO-DAC, Didier BOURIGAULT, Frédéric BILHAUT.

Références

AURNAGUE M., BRAS M., VIEU L. et ASHER N. (2001). « The Syntax and Semantics of Locating Adverbials ». In *Cahiers de Grammaire*, 26, 11-35.

BERRI J., CARTIER E., DESCLÈS J.-P., JACKIEWICZ A. et MINEL J.-L. (1996). « Filtrage automatique de textes ». In *Natural Language Processing and Industrial Applications*. Moncton, Canada.

BIBER D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.

BIBER D. (1989). « A typology of english texts ». In *Linguistics*, 27, 3-43.

CHAROLLES M. (1997). « L'Encadrement du Discours, Univers, Champs, Domaine et Espaces ». In *Cahiers de Recherche linguistique*, 6.

CRISPINO G., HAZEZ S. B. et MINEL J.-L. (1999). « Architecture logicielle de ContextO, plate-forme d'ingénierie linguistique ». In P. Amsili (éd.), *Actes de TALN 1999 (Traitement automatique des langues naturelles)*. TALN, Cargèse.

EDMUNDSON H. (1969). *New methods in automatic abstracting*. Journal of ACM.

FERRARI S., BILHAUT F., WIDLÖCHER A. et LAIGNELET M. (2005). « Une plate-forme logicielle et une démarche pour la validation de ressources linguistiques sur corpus : application à l'évaluation de la détection automatique de cadres temporels ». In *Actes des 4èmes Journées de Linguistique de Corpus*. Lorient.

GROSZ J. et SIDNER A. (1986). « Attention, intentions, and the structure of discourse ». In *Computational linguistics*, 3 (12).

HO-DAC M., JACQUES M.-P. et REBEYROLLES J. (2004). « Sur la fonction discursive des titres ». In S. Porhiel et D. Klingler (éds.), *L'unité texte, Actes du colloque 'Regards croisés sur l'unité texte / Conjoint Perspectives on Text'* : Perspectives. Chypre.

MARCU D. (2000). « The Rhetorical Parsing of Unrestricted Texts : A Surface-Based Approach ». In *Computational Linguistics*, 26.

WIDLÖCHER A. et BILHAUT F. (2005). « La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus ». In M. Jardino (éd.), *Actes de la 12e Conférence Traitement Automatique du Langage Naturel (TALN)* : LIMSI. ATALA, Dourdan, France.