Fouille de textes hiérarchisée, appliquée à la détection de fautes

Nadine Lucas et Bruno Crémilleux

GREYC CNRS UMR 6072
Université de Caen
Boulevard du Maréchal Juin
F-14032 Caen cedex

Nadine.Lucas@info.unicaen.fr, Bruno.Cremilleux@info.unicaen.fr

RÉSUMÉ *Cet article présente une approche hybride de fouille de données textuelles basée sur des segments de textes basés sur la hiérarchie de mise en forme. Elle fait coopérer des outils de fouille de données, des outils de structuration et d'analyse robustes de document et la linguistique de discours. L'application concerne la détection de l'absence et la présence de fautes de style dans des articles scientifiques en anglais. Nous décrivons d'abord les unités textuelles basées sur la hiérarchie du texte. Leurs descripteurs combinent des informations sur les formes et les positions relatives des marqueurs stylistiques. La méthode met en avant la nécessité de garder le contexte étendu d'un marqueur textuel. Nous présentons ensuite deux méthodes de fouille mises en œuvre pour caractériser la correction en anglais : règles de caractérisation et motifs émergents. Enfin, nous évaluons les résultats obtenus par l'application de ces techniques.*

ABSTRACT. *This paper presents an original text mining approach based on text segmented units, mixing data mining techniques and text linguistics. First, we describe the text units and their descriptors designed for detecting mistakes in scientific papers in English. These include text organisation, emphasizing the relative positions and the context of stylistic markers. Secondly, the paper explains the text mining methods required for such tasks and provides the techniques to extract rules characterizing classes and emerging patterns techniques. Experiment results show the usefulness of text organisation set as an hypothesis. The paper concludes on the text mining benefits for linguistic tasks.*

MOTS-CLÉS : *fouille de textes, linguistique textuelle, mise en forme matérielle, structure typographique, stylistique, règles d'association, motifs émergents, caractérisation de classes, détection de fautes d'anglais, correction, style.*

KEYWORDS: *text mining, textual linguistics, stylistics, document layout, text structure, association rules, emergent patterns, emerging patterns, class characterization, detection of English mistakes, style errors.*

Text mining based on document layout: application to check style

Document numérique. Volume 8 – n° 3/2004, pages 107 à 133

1. Introduction

Nous exposons ici un jeu d'essais préliminaires destinés à valider une idée simple : l'organisation ou la « disposition » des documents mérite d'être prise en compte dans la fouille de données textuelles. Le changement de support des documents, qui sont désormais de plus en plus des documents électroniques accessibles par Internet, a entraîné des évolutions techniques dont les langages d'annotation de type XML. Cependant, il existe toujours un vide entre les moyens techniques disponibles pour assurer un « rendu » typographique et les moyens d'appréhender automatiquement la logique qui sous-tend leur usage. La tendance actuelle est de chercher à normaliser les feuilles de style ou DTD[1] pour qu'elles deviennent univoques et informatiquement exploitables à peu de frais. Pourtant, les utilisateurs ne se plient guère aux recommandations. Parallèlement, l'exploitation de données disponibles, notamment la structure hiérarchique des documents est restée sinon ignorée, du moins peu exploitée en fouille de textes.

Nous distinguons ici la mise en forme matérielle des textes, telle qu'elle est observable (Chali et coll., 1996), et l'exploitation de cette organisation, désignée comme disposition. Si l'idée est partagée par beaucoup, sa mise en œuvre se heurte à de nombreuses difficultés (Hearst 1994 ; Pascual et Péry-Woodley, 1997). Nous partons d'une application particulière, mais nous avons cherché à cerner les principaux problèmes qui peuvent se poser dans la prise en compte de la disposition. Nous utilisons des techniques robustes de structuration de documents et d'annotation via les marqueurs de surface du texte (voir la section 2.3). Les connaissances conservées sont utiles dans la perspective de la fouille de données pour les cas difficiles, phénomènes rares, données denses, que l'on rencontre fréquemment dans la fouille de textes. Les résultats obtenus en exploitant *explicitement* certains niveaux de la hiérarchie du texte par la méthode de recherche d'associations et de caractérisation de classes d'une part, par la méthode des motifs émergents d'autre part, nous encouragent à persévérer dans cette voie. Nous relatons cette expérience, certes limitée, car elle ouvre des perspectives plus riches que celles que nous escomptions.

Nous nous intéressons à une activité minutieuse et coûteuse appelée correction éditoriale ou *copyediting*, par opposition à la correction scientifique assurée par les comités de lecture. Nombre d'éditeurs souhaitent passer à la correction semi-automatique des fautes de ponctuation, de grammaire ou de style relevant du *copyediting*. Le besoin industriel est l'élément déclencheur de nos travaux sur la révision d'articles scientifiques internationaux rédigés en anglais. Nous cherchons en particulier à détecter automatiquement la présence ou l'absence de fautes dans ces articles, ou peut-être devrions-nous dire la conformité ou la non-conformité aux normes de style établies par la maison d'édition. Nous verrons qu'une telle tâche

[1] Document Template Data ou description du formatage des documents

requiert une coopération de techniques de fouille de données et de connaissances linguistiques et stylistiques.

Avant d'aborder ces questions, remarquons que les techniques les plus courantes de fouille de données textuelles reposent sur une description élémentaire des textes basée sur les mots, même si les techniques subséquentes sont élaborées. Citons, par exemple, l'exploitation des vecteurs obtenus après l'utilisation d'un thésaurus et/ou l'élimination de mots vides de sens (Salton, 1989). Les textes sont alors vus comme des « sacs de mots ». Certaines techniques s'appuient sur la phrase, ainsi, le marquage de termes par des étiquettes provenant du traitement automatique des langues (« tagging ») (Feldman et coll., 1998 ; Riloff et Lehnert, 1994). Les textes sont alors vus comme des « tas de phrases ». Citons aussi les travaux sur les termes (associations pertinentes de mots) plus élaborés que ceux qui sont basés sur les mots (Roche et Kodratoff, 2003 ; Jacquemin et Bourigault, 2003). À la différence de ces approches, qui ne tiennent pas compte de la mise en forme matérielle, ou de l'organisation du document en tant que telle, nous proposons dans cet article l'utilisation de marqueurs linguistiques ayant l'originalité d'être caractérisés par le niveau de la hiérarchie textuelle concerné. Les textes sont alors vus alternativement comme des « tas de parties », « tas de paragraphes », « tas de phrases ». Nous avons ajouté une mesure supplémentaire, le virgulot,[2] chaîne de caractères délimitée par une virgule. Nous verrons en quoi les descripteurs associés à ces segments de texte sont appropriés à la détection de la conformité dans des articles scientifiques en anglais.

La pertinence des descripteurs et des unités manipulées est validée par l'exploration d'un corpus d'apprentissage de 18 paires d'articles avant et après révision éditoriale, au moyen de deux méthodes de fouille de données. Nous avons procédé d'une part à la production de règles de caractérisation, d'autre part à la recherche de motifs émergents. Les données à traiter sont quantitativement importantes, puisque nous effectuons la fouille dans différentes mesures de texte à savoir la partie, le paragraphe, la phrase et le virgulot. Ce découpage est basé sur la notion linguistique de « sélection » justifiée méthodologiquement. Il correspond, pour notre jeu de données initiales, à 728 parties, 3 520 paragraphes, 10 643 phrases et 22 041 virgulots.

La contribution de ce travail réside dans l'utilisation explicite de l'organisation du texte, par l'élaboration de marqueurs textuels sophistiqués, intégrés en tant que descripteurs à des techniques de fouille de textes performantes. L'ensemble permet à la fois de valider des hypothèses linguistiques et d'évaluer les procédures mises en œuvre, tant dans le pré-traitement que dans la recherche de régularités. *In fine*, elles ont permis de mettre en lumière certains traits caractérisant l'absence de fautes de grammaire et de style dans des textes scientifiques.

[2] Néologisme définissant une division interne à la phrase, délimitée par la virgule.

La section 2 présente brièvement l'état de l'art, puis l'application visée et les hypothèses linguistiques conduisant à l'élaboration des descripteurs sur lesquels repose la caractérisation des textes. La section 3 détaille le pré-traitement. La validation de nos hypothèses nécessite une recherche efficace de co-occurrences fréquentes entre descripteurs. Les techniques de fouille de données textuelles mises en jeu, par extraction d'associations et par motifs émergents sont décrites à la section 4. La section 5 donne les résultats expérimentaux et discute la validité des hypothèses émises. Enfin, la section 6 établit un bilan des expériences et dresse quelques perspectives pour étendre la portée de ce travail.

2. Analyse du problème

2.1 Fouille de données et de textes

La fouille automatique de textes concerne les processus interactifs et itératifs de découverte de connaissances dans de grandes collections de documents (Feldman, 1995 ; Kodratoff, 1999). La fouille de textes est définie par Sebastiani comme l'ensemble des tâches qui, par analyse de grandes quantités de textes et la détection de modèles fréquents, essaie d'extraire de l'information probablement utile (Sebastiani, 2002). Fouille de données et fouille de textes possèdent en commun des méthodes et algorithmes tels que les algorithmes de recherche par niveaux de motifs vérifiant certaines propriétés (Mannila et Toivonen, 1997), l'exemple le plus classique de propriété étant la fréquence.

La fouille de textes présente ses propres spécificités : les documents peuvent être plus ou moins structurés, la phase de pré-traitement joue un grand rôle. Il est courant et souvent intéressant de prendre en compte l'ordre des mots, et il est maintenant établi que des ressources provenant du traitement automatique des langues et/ou de la linguistique sont nécessaires pour apporter des résultats applicatifs avec une réelle plus-value. Beaucoup de méthodes de fouille de textes incluent des traitements d'analyse pré-lexicale (e.g., traitement des chiffres), d'analyse lexicale telle que l'élimination de « mots vides », morphologie, analyse syntaxique (e.g., détermination des groupes nominaux), analyse sémantique incluant des apports linguistiques aussi bien que des particularités des textes étudiés.

Du côté de l'analyse linguistique, l'approche du texte par des marqueurs de surface, détaillée en section 2.3, est connue chez les linguistes de corpus. Elle est exploitée entre autres par Péry-Woodley qui a travaillé sur un corpus académique anglais langue maternelle ou langue seconde (Péry-Woodley, 1991). Ses travaux portent également sur la relation entre marques linguistiques et mise en forme matérielle (Pascual et Péry-Woodley, 1997). On pourra consulter d'autres ouvrages sur le genre académique (Fløttum et Rastier, 2002) sur l'anglais académique (Biber, 1988 ; Swales, 1990) ou sur l'anglais académique, langue seconde (Fontaine et Kodratoff, 2003).

Cependant, dans les analyses de discours contemporaines, et spécialement en informatique linguistique, l'analyse est implicitement limitée au paragraphe (Marcu, 2000). L'approche liant stylistique et extraction d'information automatique a été expérimentée sur un corpus internet, principalement journalistique, étudié au niveau de la phrase (Kando, 1999 ; Karlgren, 2000). Les travaux de Hearst portent sur des groupes de paragraphes (Hearst, 1994). La structure HTML du document est explicitement exploitée pour la recherche d'informations orientée Internet (voir par exemple Muslea et coll., 2003) ou pour la génération de résumé (Kiyota et Kurohashi, 2001), mais non la structure linguistique. Notre originalité consiste à prendre en compte différents niveaux d'analyse du texte : de la partie au virgulot, et ce à la fois du point de vue linguistique et du point de vue de l'organisation matérielle.

2.2. Usage multiple des représentations condensées de motifs

Une tâche classique en fouille de textes est de trouver des motifs fréquents ou caractérisant des classes (un motif est un ensemble de descripteurs de textes). Une des originalités de notre approche est de situer ce travail dans le cadre plus général des usages multiples des représentations condensées. Cet article n'a pas pour but de présenter ce cadre. Néanmoins, nous pouvons donner une définition intuitive d'une représentation condensée et indiquer ses usages. Une représentation condensée de motifs est une synthèse des données, mettant en évidence certaines propriétés (par exemple, les motifs vérifiant la classique propriété de fréquence). On parle de représentation car il est possible de regénérer tous les motifs vérifiant les propriétes et elle est condensée car elle contient bien moins de motifs que le total. Une définition plus formelle et un cadre général sont donnés dans (Calders et Goethals, 2002). Il existe plusieurs sortes de représentations condensées de motifs, les plus courantes étant les représentations condensées à base de motifs fermés, libres (ou clés) ou encore les δ-libres.

L'utilisation des représentations condensées présente un double avantage. Premièrement, elle permet d'améliorer l'efficacité des algorithmes pour des tâches usuelles comme la découverte de règles d'association. Même si cette technique est aujourd'hui bien maîtrisée, l'utilisation de représentations condensées permet de réussir l'extraction des règles dans des contextes où les algorithmes comme Apriori échouent (Bastide et coll., 2002 ; Boulicaut et coll., 2003). Historiquement, c'est cette motivation qui a suscité l'introduction des représentations condensées. Deuxièmement, les représentations condensées autorisent des usages multiples des motifs et sont un solide point de départ pour la réalisation de différentes tâches d'exploration de données : les classiques règles d'association (Agrawal et coll., 1993), celles extraites de textes (Cherfi et coll., 2003), les règles informatives ou non redondantes (Zaki, 2000), les règles à prémisses minimales utilisées par exemple pour la caractérisation de classes (Crémilleux et Boulicaut, 2002), les règles de classification (Liu et coll., 2000), le clustering à base d'association ainsi que la

production de concepts donnant une définition en intension et en extension d'un cluster (Han et coll., 1997 ; Durand et Crémilleux, 2002), la fréquence d'une expression booléenne quelconque (Manilla et Toivonen, 1996). Les représentations condensées sont un point clé dans de nombreuses applications pratiques. Ces méthodes d'analyse sont aujourd'hui plébiscitées par les experts, qui savent qu'elles permettent de dégager une large plus-value sur leurs données.

Dans ce travail, nous cherchons, au final, à caractériser des segments d'articles à corriger ou pas (cf. le paragraphe suivant). Aussi, nous nous intéressons d'une part aux règles de caractérisation afin de faire ressortir ce qui est intrinsèque à chaque classe puis aux motifs émergents pour mettre en évidence les motifs séparant au mieux les classes. Le cœur de la méthode des règles de caractérisation est donnée dans (Crémilleux et Boulicaut, 2002) et celle des motifs émergents issus de représentation condensée dans (Soulet et coll., 2004).

Au-delà de l'élégance de la démarche, le cadre des représentations condensées favorise la factorisation de tâches du pré-traitement. Ainsi, la préparation des données (cf. paragraphe 3) est unique que ce soit pour les règles de caractérisation ou pour les motifs émergents.

2.3. *Une application, la révision éditoriale*

La révision éditoriale, aussi appelée correction éditoriale, ou *copyediting* prend place après la correction scientifique assurée par les comités de lecture, et juste avant l'impression des revues. Les articles ont donc déjà été acceptés sur le fond et la forme. La révision est actuellement assurée par des relecteurs anglophones pour garantir une certaine unité dans les publications, un certain niveau de qualité et de *lisibilité* des articles. La conformité aux normes de mise en page fait partie du problème, mais aussi, et c'est ce qui nous intéresse ici, le niveau soutenu d'anglais, ce qui inclut la correction des fautes de ponctuation ou de style. Il est important de noter dès l'abord que les modifications sont généralement peu nombreuses. L'examen des corrections apportées aux manuscrits montre que les déplacements de ponctuation, notamment des virgules, constitue le plus gros pourcentage de modifications. En moyenne, on rencontre 17% de segments modifiés sur notre échantillon. Cependant, la fourchette établie sur un corpus plus étendu est large : de 0 modification à 85 modifications par article. Celles-ci concernent, outre la ponctuation, l'ordre des mots (en particulier la place des adverbes), le choix de l'article défini ou indéfini, le choix du temps ou de la voix, le choix des conjonctions, des tournures etc. Il n'est donc pas exagéré non plus de parler de correction de fautes. Un autre aspect de la révision est que les remaniements peuvent concerner des passages de texte étendus. On trouve des phrases déplacées ou ré-écrites, voire des paragraphes déplacés ou profondément modifiés.

Notre objectif à moyen terme est de classer automatiquement les articles selon leur qualité et dans un premier temps en deux catégories : ceux qui ont besoin d'être

retouchés ou corrigés, et ceux qui n'ont pas à être modifiés, autrement dit, théoriquement nous cherchons à mettre de côté les articles similaires à ceux qui présentent 0 modification pour tous les niveaux de segmentation. En pratique, on obtient des caractérisations pour chaque mesure. Nous verrons (cf section 5) que les caractérisations sont meilleures pour les segments sans faute.

2.4. *Hypothèses*

Une étape d'acquisition de connaissances est nécessaire pour caractériser les textes. Cette étape est fondée sur la comparaison des textes dans les deux versions, avant et après révision humaine. Nous disposions de 18 paires d'articles et chaque article comptait peu de modifications. Pour résoudre ce problème de données rares nous avons choisi d'exploiter l'organisation du texte en priorité, puisqu'il paraissait plus raisonnable, à la fois conceptuellement et techniquement, de procéder par paliers et de morceler la difficulté. La structure hiérarchique des documents est assez stable et bien reflétée par la mise en forme matérielle dans le corpus. Les parties sont comparables aux parties, les phrases aux phrases etc.

Le second choix consiste à orienter la fouille. Alors que le choix des marqueurs est déterminant, la question des critères de choix des descripteurs est très peu abordée dans la littérature (Daille et coll., 1996). On peut espérer caractériser les fautes et construire un jeu de descripteurs décrivant des erreurs de style ou de grammaire ou alors caractériser les cas normaux. Quelle que soit l'orientation adoptée, nous devons définir un contraste. Nous avons choisi de caractériser la correction ou conformité aux normes. Nous observons que les textes sont le plus souvent corrects, et nous supposons aussi que nous pouvons caractériser la correction stylistique (comme cohésion et cohérence, voir ci-dessous). Cette hypothèse est fondée sur l'observation d'un corpus où la variété des fautes est très grande, ce qui empêche la définition de fautes typiques. Cependant, la définition de la norme pose des problèmes dans la mesure où les réviseurs ne corrigent pas de façon uniforme : la norme idéale ne coïncide pas exactement avec la norme statistique, ce qui introduit une difficulté supplémentaire.

Choix des unités à caractériser

La mise en forme matérielle ou organisation textuelle est une source essentielle d'information. En effet, notre problème peut se reformuler pragmatiquement comme suit : comment isoler les textes bien écrits ou du moins suffisamment bien écrits, donc non modifiés par le réviseur humain? Quelles sont les unités textuelles caractérisables ? Lorsque nous avons évoqué les modifications des textes, nous avons indiqué que les déplacements venaient largement en tête. Il s'ensuit que les mots ne sont pas des descripteurs plausibles, comme ils peuvent l'être dans les études fondées sur le lexique, pour des tâches d'indexation, par exemple (Bernard et coll., 2002). On peut également utiliser des étiquettes grammaticales (Roche et Kodratoff, 2003),

mais comme celles-ci sont établies au niveau de la phrase, elles ne permettent pas de traiter notre problème. Rappelons en effet que les fautes de grammaire sont peu nombreuses. Les fautes de compréhension et les fautes de style, les plus nombreuses dans notre corpus, concernent des phrases jugées syntaxiquement correctes par des anglophones lorsqu'elles sont présentées hors-contexte. Les modifications n'en sont pas moins jugées pertinentes en contexte, car elles améliorent la lisibilité.

Nous avons choisi de travailler non sur des articles, mais sur les subdivisions de l'article, en nous appuyant sur les divisions les plus stables. Nos *mesures* sont les parties titrées, les paragraphes, les phrases, les virgulots (ou chaînes de caractères délimitées par une virgule). Les espaces de recherche ainsi définis ont statut d'unités textuelles caractérisables. Cela permet en effet d'approximer des zones de portée permettant de traiter séparément les différents types de fautes (de grammaire, de compréhension et de style). Nous rechercherons dans ces espaces des chaînes de caractères ou des caractères, contigus ou disjoints.

L'avantage de cette approche tient bien sûr à la division de la difficulté. Nous multiplions aussi les données à traiter, puisque chaque texte initial sera divisé autant de fois qu'il existe de mesures ou niveaux sur notre échelle. C'est ainsi que les 36 textes (18 paires) équivalent à 728 parties, 3 520 paragraphes, 10 643 phrases et 22 041 virgulots, appelés segments de texte. Il est également plus commode de gérer le contexte, comme on le verra, une fois que les mesures sont clairement définies sur une échelle de niveaux et objectivées.

Choix des descripteurs

Le choix d'unités textuelles caractérisables entraîne le choix des descripteurs devant justement caractériser les segments de texte. Il n'est pas du ressort de cet article d'expliquer précisément sur quelles bases les descripteurs sont établis. Les formes choisies correspondent aux difficultés de l'expression écrite en anglais langue seconde (Péry-Woodley, 1991). Nous proposons ici un aperçu de la méthode de structuration textuelle (Lucas et coll., 1993) appliquée au problème posé (Turmel et coll., 2003).

Les hypothèses linguistiques mises en jeu dans notre travail tiennent en deux points importants. Le premier est que la correction du style peut s'observer sur chaque segment pris isolément. Cette idée s'appuie sur la notion de sélection (Harris, 1952), la sélection désignant une fenêtre d'observation des phénomènes linguistiques, qui recouvre les mesures de texte définies plus haut. On reconnaît en effet plusieurs types de fautes associées à des portées différentes. Les fautes de grammaire s'observent dans la phrase, les fautes de compréhension dans le paragraphe ou la section, les fautes de style mobilisent des contextes plus étendus. On devrait donc observer la présence récurrente de formes remarquables à des positions remarquables, caractérisant des segments, dans une sélection ou mesure. Le second point concerne la prise en compte de la hiérarchie textuelle, ou du contexte nécessaire à la compréhension, suivant en cela les travaux stylistiques de l'école russe (Bakhtine,

1984). Alors qu'il est habituel de séparer les niveaux d'analyse de la langue par rapport au discours, à la phrase, au mot, nous avons choisi de combiner ces niveaux.

La première de nos hypothèses linguistiques fortes est d'appliquer la notion de sélection à *toutes* les mesures du texte (Lucas et coll., 1993). Rappelons que la dissociation de la forme et de la position est une précaution méthodologique préconisée par les distributionnalistes[3]. La position début ou fin est relative à la mesure ou sélection considérée. Les segments des articles académiques en anglais sont ainsi décrits par des patrons corrects, établis pour différentes portées de macro-constructions textuelles. Les formes remarquables utilisées comme indices formels sont ici appelées « marqueurs ». Ces marqueurs sont organisés en classes (abstraites) en fonction de leur portée et mis en correspondance avec une sélection (concrète) (cf. tableau 1). Les marqueurs sont des lettres, par exemple *ed/ought*, des mots comme *and*, des groupes de mots ou encore des expressions plus complexes, comme les appels à référence bibliographique. En règle générale, les marqueurs peu fréquents et longs ont une portée plus grande que les mots fréquents et courts.

Tableau 1. *Exemples de marqueurs*

Sélection	Formes remarquables		
Virgulots	*-ly* *-ed* *Its / Their* *-ing* …	*and* *despite* *such* *indeed* *because …*	: ; "" ( ) ( )
Phrases	Idem + majuscule ou point		
Paragraphes	*There is* *In fact* *As well* …	Déterminants indéfinis Adverbes Pronom Conjonction subordination	Date Référence biblio
Sections		Conjonctions majeures Aspect Voix …	Chiffres Ordinal
Parties	*in spite of* *for this reason* *as well as*	Connecteurs adverbiaux Personnel / Impersonnel Futur /passé Anaphore Conjonctions *wh …*	

Il est assez aisé de comprendre que les lettres *ed/ought* ne caractérisent le preterit qu'à la condition qu'elles soient en finale de mot, d'où l'importance de la position. Nous établissons une corrélation entre la sélection et les marqueurs appropriés à cette

[3] Distributionnel a aussi un sens statistique, ici nous nous référons à une pratique propre à la linguistique

sélection. Les terminaisons de mot comme *ed/ought* (marque du temps du verbe, ici le preterit) caractérisent une proposition, unité linguistique abstraite contenant un seul verbe conjugué dans sa définition classique. Une proposition peut s'inscrire dans un virgulot, ou une phrase, unités typographiques plus accessibles lors d'un traitement informatique. De la même façon, une phrase pronominale (apparaissant en fin de paragraphe) est caractérisée par *It*, un paragraphe en coordination disjonctive (indiquant une opposition en fin de partie) par *By contrast*. Nous caractérisons ainsi tous les segments par présence ou absence de marqueurs détectés à une position remarquable (voir section 3. 2). Dans l'implémentation, les sections et parties ne sont pas distinguées, la mesure « partie titrée » est donc associée aux marqueurs de deux sélections, section et partie.

S'il est banal que des mots, termes ou patrons soient utilisés comme descripteurs en fouille de textes, il est encore inusité de prendre la position comme critère de description. Les descripteurs issus d'un étiqueteur de Brill contiennent une information latente sur l'ordre des mots dans la phrase (Brill, 1994). Dans notre approche, la position (début/fin) à tous les niveaux est le critère supplémentaire qui permet de passer de la forme (le marqueur) au « descripteur » proprement dit. En effet, les descripteurs incluent des informations explicites sur la position, et ils sont toujours associés à une mesure.

L'importance du contexte

Notre seconde hypothèse linguistique forte est qu'il existe une relation entre les descripteurs des différents niveaux depuis les plus hauts. Elle s'appuie sur des considérations de rhétorique permettant de caractériser l'unité de style (Enkvist, 1985). Certains linguistes du discours font appel aux notions de *cohésion* et de *cohérence* (Tierney et Mosenthal, 1981 ; Parsons, 1990). Dans la suite de l'exposé, nous parlerons de « facteurs de cohésion » pour les descripteurs associés à une sélection prise isolément, et de « facteurs de cohérence » pour ceux qui reflètent les informations héritées du contexte. Mais la difficulté tient à l'expression technique de la notion de cohérence. Nous l'avons traduite en terme d'héritage entre niveaux, notion plus familière en informatique.

La structure textuelle est une *hiérarchie inclusive*. Une fin de phrase peut correspondre également à une fin de paragraphe, et à une fin de partie. Cette notion implique la mise en mémoire du contexte, à différents grains. Chaque segment « hérite » en quelque sorte des descripteurs positionnels et formels qui caractérisent la mesure englobante. Ainsi, une phrase « connaît » non seulement ses descripteurs, mais aussi ceux de la mesure paragraphe et ceux de la mesure partie dans lesquelles cette phrase est située. C'est ce qui permet de conserver la trace du contexte, malgré l'éclatement des documents en différentes mesures.

La représentation que nous avons adoptée diffère ainsi considérablement des représentations courantes, basées sur des réseaux sémantiques lexicaux, sur des étiquettes catégorielles ou encore sur l'extraction de chaînes de caractères répétés

(Chan et coll., 2002). Elle se fonde en effet sur une grammaire de texte et de discours, laquelle exploite des indices morphologiques et la position relative de ces indices, facteurs de cohésion, ainsi que le niveau hiérarchique, facteur de cohérence, ce qui est essentiel dans le cas qui nous préoccupe (Lucas et coll., 2003). Les descripteurs sont donc complexes.

Dans la section suivante nous exposons plus en détail les traitements nécessaires pour l'implémentation.

3. Mise en œuvre

Nous avons opté pour la caractérisation de segments de texte qui peuvent être estampillés comme invariants (non modifiés) ou variants (modifiés) dans la comparaison des versions avant et après révision humaine. Nous avons donc besoin d'un autre descripteur, issu de la comparaison des deux versions d'un même article, avant et après révision. Ce descripteur sera appelé classe par la suite. Nous estampillons « faute » dans la version avant révision un segment de texte qui a été modifié dans la version corrigée, et « non-faute » un segment qui n'a pas été modifié ultérieurement.

La première étape nécessite la conversion du format des documents électroniques. Le corpus se compose d'articles au format *RTF* (Rich Text Format) et d'autres au format Microsoft® Word (.doc). Nous avons converti tout le corpus au format HTML et XML. La seconde étape nous conduit à un découpage du document en différents segments typographiques appartenant à une hiérarchie de mesures : les parties titrées, les paragraphes, les phrases, les virgulots. Nous utilisons un segmenteur robuste de document élaboré au GREYC et écrit en Perl pour mener cette tâche à bien (Voisin, 2002). Notons que dans ce travail, les sections et sous-sections n'ont pas été prises en compte en tant que telles, elles sont comptées comme partie, la mesure supérieure au paragraphe. Il s'agit là d'une simplification.

3.1 Alignement

Dans la comparaison des versions avant et après révision humaine, les parties sont comparées aux parties, les paragraphes sont comparés aux paragraphes etc. La difficulté de notre approche tient aux fluctuations de frontières : les mesures définies ne sont pas absolument stables, en particulier les virgules sont souvent déplacées entre la version avant et la version après révision humaine. Un problème d'alignement se pose donc.

Nous avons opté pour une caractérisation descendante et dans un premier temps purement métrique des segments à aligner en nous appuyant sur le nombre des items appartenant à un groupe. Nous tirons parti de la hiérarchie inclusive des constituants textuels pour profiter de l'efficacité des traitements en cascade. Cette fois, nous

traitons toutes les unités depuis la partie jusqu'à la lettre (le caractère) qui se trouve être l'atome du traitement informatique. Ainsi, les paragraphes invariants comptent le même nombre de phrases, tandis que les paragraphes retouchés métriquement sont les paragraphes dans lesquels le nombre de phrases a changé, soit que des phrases aient été refondues, soit qu'une phrase ait été scindée et ainsi de suite.

Dans un second temps, nous testons la distance d'édition ou l'identité des segments constituants en vérifiant s'ils ont un même début et une même fin. Cette vérification permet de détecter des changements locaux appelés ajouts ou suppressions. Nous parvenons ainsi à limiter considérablement les tests de comparaison. Au terme de la procédure d'alignement, nous obtenons la mise en correspondance correcte des différentes mesures.

3.2 Marquage

Les articles du corpus subissent un pré-traitement, comme cela est fréquemment le cas dans le domaine de la fouille de textes. Ce pré-traitement se distingue des pré-traitements classiquement effectués, en vertu de l'hypothèse d'exploitation de l'organisation textuelle présentée ci-dessus. La recherche de marqueurs textuels consiste à relever dans les articles les formes qui nous intéressent, dans les espaces de recherche qui nous intéressent, à des positions qui nous intéressent. Nous détaillons ici la recherche des marqueurs et celle des descripteurs proprement dits.

La recherche de marqueurs stylistiques dans les segments d'articles est menée de façon robuste par un étiqueteur réalisé au GREYC (Voisin, 2002). Les différents segments sont traités séparément. Techniquement, nous utilisons des expressions régulières pour détecter des mots (*also, it*) ou patrons (*is/are… -ed…by*) caractérisant par exemple la voix passive. Chaque sous-espace de recherche est stipulé, toujours par l'intermédiaire du langage des expressions régulières.

Les descripteurs sont établis à travers une procédure de vérification de la position. Les marqueurs sont étiquetés suivant leur position dans chaque mesure, « début » « fin » ou « milieu ». Les positions remarquables sont début et fin, elles sont donc reportées sur le descripteur. Les critères positionnels sont ainsi associés à chaque marqueur (par défaut, la position est milieu).

Il existe un descripteur plus complexe, nommé « isopériphérie » : ce descripteur permet de noter la présence d'un même marqueur en début et en fin de segment. Par exemple, une phrase qui commence par un adverbe (le mot caractérisant une phrase) et se termine par un adverbe porte le descripteur « isopériphérie ». De même, une partie qui commence et se termine par une phrase passive (la phrase caractérisant une partie) porte le descripteur « isopériphérie ».

Enfin, pour représenter l'héritage du contexte, les descripteurs sont annotés par niveau : par exemple, un virgulot « connaît » non seulement ses descripteurs notés

NV mais aussi ceux de sa phrase notés NF, ceux de son paragraphe notés NPA et ceux de sa partie notés NP.

4. Caractérisation des segments avec ou sans faute

Nous présentons succinctement dans cette section les méthodes de fouille de données (règles de caractérisation de classes et de motifs émergents) nécessaires à la validation des hypothèses énoncées à la section 2 et à la compréhension de l'ensemble de ce travail. Rappelons que dans notre travail la fouille s'effectue suivant les différentes *mesures* du texte (partie titrée, paragraphe, phrase, virgulot). Nous commençons par présenter les notations nécessaires.

4.1. Représentations condensées : notations

Les méthodes de fouille de données sont classiquement présentées dans un contexte transactionnel et nous commençons par définir les notions d'item et de motifs. Dans ce qui suit, les définitions sont sous-entendues relatives à un ensemble d'*exemples* (par exemple, des virgulots ou des phrases caractérisés par leurs descripteurs).

Définition 1 (item, motif) *Soit D = {D_1,..., Dn}, un ensemble de descripteurs. Un élément de D est appelé item et un sous-ensemble de D motif (ou itemset).*

Dans notre contexte, un item est par exemple la présence de la finale *–ing* dans un virgulot. Un motif (i.e. itemset) est par exemple *or to ing parentheses.* Les motifs traduisent des associations entre descripteurs. D'un point de vue technique, pour des raisons d'efficacité, les algorithmes travaillent à partir d'items binaires. Notons qu'un ensemble de descripteurs peut toujours être transformé en items binaires et le jeu de descripteurs que nous utilisons (cf. tableau 1) est ainsi codé avec 270 items. Nous introduisons maintenant la notion de motif fréquent.

Définition 2 (motif fréquent) *Soit F(X) (ou fréquence de X) le nombre d'exemples de la collection possédant chaque item de X. Soit* γ *un seuil de fréquence inférieur ou égal au nombre d'exemples. Un motif X est* fréquent *si* $F(X) \geq \gamma$.

L'extraction de tous les motifs fréquents est au cœur de nombreuses représentations condensées (Calders et Goethals, 2002), même s'il est possible de concevoir des représentations condensées basées sur des contraintes autres que la fréquence. La recherche des motifs fréquents dans de grands jeux de données est un problème algorithmiquement difficile car la complexité des algorithmes d'extraction de *tous* les motifs fréquents est exponentielle suivant le nombre de descripteurs. De nombreuses recherches (Boulicaut et coll., 2000 ; Bastide et coll., 2002) visent à

identifier des situations pratiques pour lesquelles les calculs restent faisables, quitte à faire des concessions sur l'exactitude des mesures d'intérêt ou bien sur la complétude des extractions. En ce qui nous concerne, nous avons utilisé une représentation condensée à base de motifs δ--libres pour les règles de caractérisation de classes et à base de motifs fermés pour les motifs émergents. Ces choix sont explicités dans les paragraphes suivants.

D'un point de vue technique, nos expérimentations ont été réalisées avec l'extracteur MVMiner développé par F. Rioult au GREYC. Etant donnés des seuils δ et γ, MVMiner produit tous les motifs δ-libres fréquents. A partir de motifs δ-libres particuliers (les 0-libres), il est aisé d'obtenir la représentation condensée des motifs fermés. Notons que la démarche inverse est plus complexe et fait l'objet d'actives recherches. La recherche des règles de caractérisation de classes et motifs émergents s'effectue par post-traitement de la représentation condensée : un filtrage pour les règles de caractérisation (Crémilleux et coll., 2002), un calcul de taux de croissance à partir de certains fermés pour les motifs émergents forts (Soulet et coll., 2004).

4.2. Règles de caractérisation de classes

Considérons un ensemble d'exemples où chaque exemple est étiqueté par une valeur de classe. Dans ce travail, nous avons deux classes (présence versus absence de fautes). Nous supposons aussi (ce qui est la situation la plus courante) que chaque exemple possède une et une seule valeur de la classe. Une règle δ-forte de caractérisation de classes conclut sur une des valeurs de la classe de façon assez « certaine » (cette notion sera précisée à la définition 4). Notons qu'une telle règle est un cas particulier des classiques règles d'association (Agrawal et coll., 1993).

Définition 3 (règle δ-forte de caractérisation de classes) Une règle δ-forte de caractérisation de classes est une règle à prémisse minimale de la forme X B, où X D et B D \ X (X est un motif), qui accepte au plus δ exceptions et qui conclut sur une des valeurs de la classe.

Par exemple, *or to ing parentheses faute* est une règle indiquant que si les descripteurs *or to ing parentheses* sont rencontrés, alors on est en présence d'un segment avec fautes. *X* est aussi appelé la prémisse de la règle. Les mesures usuelles de *fréquence* et de *confiance* (Agrawal et coll., 1993) expriment la sémantique associée à la « représentativité » et la « force » de l'association.

Définition 4 (fréquence, confiance, support) *Soit X* D. *La fréquence de X B est définie comme* F(*X* {*B*}) *et sa confiance est* F(*X* {*B*}) / F(*X*). *La fréquence est aussi communément appelée* support.

Les définitions 3 et 4 montrent qu'une règle δ-forte de caractérisation de classes n'est contredite que par au plus δ exemples. En d'autres termes, elle a une confiance au moins égale à 1 – (δ / γ).

Classiquement, on est intéressé par la recherche de toutes les règles « suffisamment » fréquentes et valides (i.e. des règles dont les fréquences et les confiances sont supérieures à des seuils fixés par l'utilisateur). La difficulté de cette recherche réside en fait dans l'extraction de tous les motifs fréquents, donc l'extraction de la représentation condensée (voir section 4.1). Il est hors du champ de cet article d'expliquer précisément comment de telles règles sont construites (Crémilleux et Boulicaut, 2002) et pourquoi nous utilisons la représentation condensée à base de motifs δ-libres (Boulicaut et coll., 2003). Disons simplement que les deux points importants à retenir sont l'efficacité pour l'extraction et le caractère de minimalité des motifs δ-libres.

Pour le premier point, la propriété de « liberté » vérifiée par les motifs δ-libres autorise un critère d'élagage sûr dans l'espace de recherche des motifs. Cela signifie concrètement que nous sommes capables de concevoir des algorithmes efficaces même dans le cas de grands jeux de données denses et/ou fortement corrélées. Il est ainsi possible d'obtenir des extractions faisables dans des cas où les algorithmes classiques échouent pour les seuils de fréquence demandés (Bastide et coll., 2002 ; Boulicaut et coll., 2003).

En ce qui concerne le second point, un motif δ-libre est une conjonction *minimale* d'items permettant de connaître la fréquence de tout un ensemble de motifs. Comme les règles δ-fortes de caractérisation de classes sont construites en prenant un motif δ-libre comme prémisse, cela signifie que nous sommes capables de produire les règles les plus simples (par rapport à leurs prémisses) pour conclure sur une classe, l'incertitude restant contrôlée par δ. Nous pensons que cette propriété est fondamentale pour la caractérisation de classes. Non seulement elle permet d'éviter le sur-apprentissage (i.e. règles sur-spécifiées conduisant à des erreurs de classement de nouveaux exemples), mais aussi elle facilite l'explication du classement d'un exemple. De plus, au-delà du classement de nouveaux exemples, un expert est toujours intéressé à faire émerger des données les concepts généraux sous-jacents au classement. Cela lui apporte un retour sur son domaine d'expertise. Enfin, si l'extraction est effectuée suivant une certaine condition aisément vérifiée en pratique, alors un sous-ensemble de la prémisse d'une telle règle ne permet pas non plus de conclure sur une *autre* classe. Ce point est important pour éviter des conflits afin d'employer de telles règles dans un processus de classification.

4.3. *Motifs émergents*

Un motif émergent est un motif dont la fréquence varie fortement entre deux jeux de données. Dans un problème de caractérisation de classes, ces jeux correspondent aux valeurs des classes (ici, présence versus absence de fautes). Une caractéristique essentielle d'un motif émergent est son taux de croissance (*growth rate*) entre deux classes. Pour le définir, il est nécessaire de préciser dans quelle collection d'exemples une fréquence est calculée. Aussi, nous introduisons maintenant dans la notation de

la fréquence un deuxième paramètre indiquant le jeu de données considéré pour le calcul de la fréquence. Soit S la collection d'exemples, la fréquence de X dans S se note F(X,S). Jusqu'à présent, nous avons utilisé une notation simplifiée en ommettant ce deuxième paramètre, puisqu'il n'y avait pas d'ambiguïté (i.e. F(X) = F(X,S)). |...| dénote le cardinal d'un ensemble.

Définition 5 (taux de croissance) *Soit $\{S_1, S_2\}$ une partition de S et soit X D. Le taux de croissance de X dans S_1 par rapport à S_2 est $GR_1(X) = |S_2| \times F(X, S_1) / |S_1| \times F(X, S_2)$. Si $F(X, S_1) = 0$, alors $GR_1(X) = 0$ et si $F(X, S_1) \neq 0$ et $F(X, S_2) = 0$, alors $G_R1(X) = \infty$.*

Définition 6 (motif émergent) *Soit X* D et S_i un élément d'une partition de S. *X est un motif émergent de S_i par rapport à $S \setminus S_i$ si et seulement si $GR_i(X) > 1$.*

Les motifs émergents ont été à l'origine introduits par G. Dong et J. Li (Dong et Li, 1999). Ils permettent de caractériser les classes de manière quantitative et qualitative et forment une solide assise pour construire des classifieurs (Dong et coll., 2000). Dans la pratique, on s'intéresse souvent aux motifs émergents ayant les plus forts taux de croissance. Précisons qu'il est aisé de généraliser la recherche de motifs émergents pour une partition de S ayant plus de deux classes.

La recherche des motifs émergents est un problème difficile car le nombre de candidats est très élevé et le taux de croissance n'est pas une contrainte anti-monotone qui permettrait des élagages fiables dans l'espace de recherche (Mannila et Toivonen, 1997). Classiquement, la recherche des motifs émergents s'effectue par l'extraction de deux bordures (Dong et Li, 1999 ; De Raedt et Kramer, 2001). L'une correspond aux plus longs motifs fréquents dans une classe et l'autre aux plus courts motifs peu fréquents dans l'autre classe. La recherche de cette deuxième bordure est particulièrement difficile à cause de la faible valeur de la fréquence.

Dans (Soulet et coll., 2004), nous proposons une représentation condensée exacte des motifs émergents et nous définissons les motifs émergents forts. Nous montrons comment il est possible d'inférer ceux-ci à partir de la représentation condensée des motifs fermés fréquents. Un motif émergent fort possède un double avantage : d'une part, ce sont les motifs émergents possédant les plus forts taux de croissance (tout motif émergent X est inclus dans un motif émergent fort dont le taux de croissance est supérieur ou égal à celui de X), d'autre part, leur taux de croissance s'obtient sans calcul supplémentaire à partir de la représentation condensée.

Munis de ces outils efficaces de recherche de règles de caractérisation (permettant de caractériser les classes) et de motifs émergents (permettant de distinguer ce qui est caractéristique d'une classe par rapport à une autre), nous pouvons maintenant décrire le processus expérimental mis en œuvre pour valider les hypothèses énoncées à la section 2.

5. Expérimentation et résultats

5.1 Expérimentation

L'expérimentation se passe en trois temps. D'abord, il faut chercher tous les descripteurs présents dans chaque mesure de texte du corpus étudié, c'est-à-dire dans chaque virgulot, chaque phrase, chaque paragraphe, chaque partie. Ensuite, nous extrayons les motifs fréquents, i.e. les co-occurrences fréquentes de descripteurs, dans chaque mesure de texte. Enfin, nous pourrons rechercher, parmi les motifs fréquents trouvés, des règles de caractérisation ou des motifs émergents forts, sur les valeurs de classes « faute » et « non faute ».

5.2 Validation de la prise en compte du contexte

L'hypothèse du contexte est validée par le biais de deux expériences (cf. tableau 2). Tout le protocole d'expérimentation décrit auparavant est appliqué d'une part en utilisant la sauvegarde du contexte, et d'autre part sans sauvegarde du contexte, et ceci sur le même corpus.

Tableau 2. *Résultats des extractions de motifs fréquents*

	Virgulot	Phrase	Paragraphe	Partie
Nb. de mesures	22 041	10643	3 520	728
Fréquence	60	80	50	50
δ	10	10	0	0
Nb. motifs fréquents avec contexte	48 507	815	953	29
Nb. motifs fréquents sans contexte	5	8	14	29

Comme indiqué en section 3, le contexte peut être gardé dans les extractions en associant à chaque segment les descripteurs qui décrivent les mesures de texte dans lesquelles ce segment s'inscrit (par exemple, une phrase « connaît » non seulement ses descripteurs mais aussi ceux de son paragraphe et ceux de sa partie). La différence entre les extractions sur les mesures qui ne possèdent que leurs descripteurs et celles sur les mesures qui héritent du contexte est éloquente (cf. tableau 3, les deux dernières lignes) : le nombre de motifs fréquents est raisonnable en présence du contexte, et insignifiant sans contexte. L'absence d'associations constatée si l'on ignore le contexte montre qu'il serait vain de chercher des règles de caractérisation basée uniquement sur les « facteurs de cohésion » ou co-occurrences dans une mesure, sans contexte. Cela revient à dire que les « facteurs de cohérence » représentant l'importance du contexte sont primordiaux.

5.3. *Résultats par les règles de caractérisation*

Nous détaillons maintenant les résultats de caractérisation obtenus sur les virgulots, mesure qui hérite de tous les autres niveaux.

Tableau 3. *Règles de caractérisation de classes portant sur le virgulot (22 041 virgulots dans les données)*

Règles de caractérisation	Classe	Fréquence	Confiance
NP As well as NF where	Non faute	229	1
NP Passif Imperso NPA Isoperiph NF by	Non faute	837	0,96
NP Futur NPA Isoperiph NF there	Non faute	499	0,93
NP Connect Adv NPA Adv conj NPA Isoperiph NF pointvirg	Non faute	295	0,89
NF parenthese NVthe NV it NV crochet	Faute	155	0,37
NF parenthese NVan NVthe NV that	Faute	126	0,37
NF with NF we NV as NV parenthese	Faute	140	0,36
NPA Adv conj NV deuxpoint	Faute	141	0,30
NF crochet NF pointvirg NV at	Faute	132	0,30
NF ing NV after	Faute	134	0,26
NPA Adv Prep NF the NF this NV ed	Faute	129	0,26

Les règles concluant sur « non-faute », c'est-à-dire celles qui caractérisent un segment non modifié, donc correct, sont les plus sûres. Les descripteurs ont été établis à partir d'une référence idéale (marqueurs grammaticaux et de style soutenu d'anglais). La première ligne du tableau 3 (règle 1) se lit « si un virgulot appartient à une phrase (complexe) contenant *where* et appartient à une partie contenant la marque de coordination complexe *as well as*, alors il s'agit d'un segment correct ». Voici un exemple d'un tel contexte (les marqueurs sont en gras). Ici, le descripteur de phrase *where* et le descripteur de partie *as well as* se retrouvent dans un même paragraphe, ce qui n'est pas nécessairement le cas : il suffit que le descripteur de partie appartienne à la même partie (indiqué par 1 Introduction).

Exemple 2 *Contexte correct pour les virgulots selon la règle 1*
1. Introduction

Toxoplasma gondii *is an obligate intracellular parasite capable of infecting and surviving in a wide range of nucleated cells (1). The invasive tachyzoite rapidly enters a host cell […]. Once established, the parasite replicates by an unusual internal budding process* ***where*** *by two new daughter cells are formed within each mother cell (4). While much of the cytoplasm and organelles of the mother cell are incorporated into the two daughters, other parts (the apical complex and associated secretory organelles* ***as well as*** *parts of the subpellicular*

cytoskeleton) disappear. It is not known what happens to these structures, although it has been presumed that they are broken down and reabsorbed during the final stages of the budding process.

Il est intéressant de noter qu'un grand nombre de règles s'applique à un passage de texte. Ainsi, outre la troisième phrase répondant à la règle 1, la deuxième phrase du paragraphe répond à la règle 2. La première et la dernière phrases répondent à une règle d'isopériphérie pour le trait définitoire. La quatrième phrase est également jugée correcte sur une règle (non citée dans le tableau) concernant la conjonction *While*. Certaines mesures peuvent être caractérisées par plusieurs règles qui s'appliquent indépendamment (une phrase contenant *what* peut aussi être une phrase passive par exemple).

Bon nombre de règles font apparaître le descripteur « isopériphérie » au niveau du paragraphe, ce qui, rappelons-le, note la co-occurrence d'une même forme en début et fin de mesure. La ligne 2 du tableau 3 se traduit ainsi « si un virgulot appartient à une phrase contenant *by* et appartient à un paragraphe bien caractérisé en début et fin par des phrases semblables et appartient aussi à une partie contenant un passif impersonnel, alors il s'agit probablement d'un segment correct ». Ces descripteurs signalent en effet des procédés caractérisant une partie ou section « descriptive ». De même, on trouve des passages de texte « argumentatifs » qui sont décrits par la règle de la ligne 4 du tableau : « si un virgulot appartient à une phrase contenant la ponctuation point-virgule, et appartient à un paragraphe caractérisé en début et fin par des phrases contenant une conjonction majeure (du type de *Although*) et appartient aussi à une partie contenant un connecteur adverbial, alors il s'agit très probablement d'un segment correct ».

La cohérence du texte est bien captée par le principe d'héritage du contexte. On note en effet que les règles concluant sur « non faute » présentent beaucoup de descripteurs des mesures englobantes. De fait, les virgulots corrects sont caractérisés par leur *position* dans des mesures bien marquées, et non par des marqueurs formels qui leur seraient propres. C'est là un résultat important.

A l'opposé, les règles concluant sur « faute », c'est-à-dire celles qui caractérisent des segments incorrects ne présentent pas ce caractère de cohérence. Il n'y a pas ou très peu d'héritage. Par exemple la ligne 5 du tableau se lit « si un virgulot contient à la fois *the* et *it* et des crochets (contenant généralement une référence bibliographique) et appartient à une phrase contenant des parenthèses, alors il s'agit assez probablement d'un segment incorrect ».

Au vu des résultats, on constate que les descripteurs de mesures supérieures sont trop peu nombreux, il y a assez peu de motifs fréquents pour les paragraphes et très peu pour les parties malgré un seuil placé au plus bas.

Il n'y a pas de règles de caractérisation extraites au niveau des parties. Cette mesure est inopérante pour cette méthode. Il y a peu de motifs fréquents, ce qui indique que les descripteurs sont insuffisants. Il est également patent que des

informations précieuses sont perdues au niveau des parties qui n'héritent pas de la mesure supérieure. Pour les corrections dites de style, l'écart à la norme ne peut être justifié en bonne logique que dans le cadre de l'article. Or, nous n'avons pas conservé la mesure « corps d'article ». Nous ne pouvons donc représenter que les facteurs de cohésion au niveau de la partie, et non les facteurs de cohérence.

La validation des hypothèses apparaît clairement dans les règles concluant sur « non-faute » (cf. tableau 3) et les résultats soulignent l'importance des critères positionnels. Comme on peut s'y attendre, avec un jeu de descripteurs caractérisant des textes stylistiquement corrects, et un corpus généralement correct, les règles d'association concluant sur « faute » ont un taux de confiance faible et sont peu fiables.

5.4. Résultats par les motifs émergents

Les résultats obtenus par les motifs émergents complètent ceux qui sont obtenus par les caractérisations. On constate que l'objectif de séparer les classes est bien atteint pour certaines mesures seulement. Ainsi, les parties et les paragraphes sont bien discriminés en deux classes « avec faute » et « sans faute ».

Examinons par exemple les résultats pour la mesure partie titrée. Dans la classe « avec faute » les 3 motifs émergents forts, de longueur <= 3 avec un seuil de fréquence de 4, nous donnent les trois motifs suivants (0 tient pour absent, 1 pour présent, GR est le taux de croissance).

NP as well as= 0; NP-ly = 1 ; NP ISOPERIF= 1 : GR= 16
NP FUTUR= 0; NP -ly= 1 ; NP ISOPERIF= 1 : GR= 14
NP -ly = 1 ; NP PASSIFIMPERS= 0; NP ISOPERIF= 1 : GR= 17

Le commentaire peut être ainsi formulé : une partie fautive est marquée par un adverbe en *-ly* en début et fin, et par cet adverbe seulement (marqueur inapproprié car de trop bas niveau), au lieu d'être marquée par une phrase avec construction passive impersonnelle, une phrase au futur ou une conjonction majeure (qui conviendraient au niveau partie).

Les résultats sur la classe « sans faute » présentent une majorité de motifs absents. 118 motifs courts (jusqu'à 3 items) sont extraits avec un seuil de fréquence de 4 et un taux de croissance supérieur à 2. Une partie titrée est correcte, lorsque les descripteurs de niveaux supérieurs (sélections de section ou partie) ne sont pas conflictuels. Par exemple, une partie est correcte s'il n'y a pas de conflit de temps, entre futur et passé, ou encore, s'il n'y a pas de co-occurrence (fâcheuse) de passif impersonnel et de passif personnel. Selon ces règles, l'extrait cité dans l'exemple 1 est effectivement correct, nous avons vérifié que l'ensemble de la partie l'est également.

Les résultats sur la mesure paragraphe sont bien lisibles et mettent en valeur l'importance du contexte. La classe sans faute contient 216 motifs courts (jusqu'à 3

items) extraits avec un seuil de fréquence de 5 et un taux de croissance supérieur à 2. Les données héritées du contexte sont particulièrement intéressantes. Un paragraphe appartenant à une partie marquée par la voix ou l'aspect est correct, indépendamment des autres marques locales au paragraphe. La classe avec fautes présente des motifs plus longs. Elle contient 8 motifs (longueur jusqu'à 5 items) extraits avec un seuil de fréquence de 5 et un taux de croissance supérieur à 2. Ils indiquent que les paragraphes commençant par un article indéfini ont toute chance d'être fautifs, s'ils ne sont pas inscrits dans un contexte nettement argumentatif (qui autorise une phrase indéfinie en initiale).

Toutefois, les résultats sur la mesure phrase sont moins probants. La classe avec faute est peu exploitable car les phrases avec faute ne présentent pas de caractéristiques groupées. Les phrases sans faute ont une caractérisation meilleure quoique moins tranchée que celle des paragraphes. La classe sans faute contient 251 motifs (longueur jusqu'à 5 items) extraits avec un seuil de fréquence de 100 et un taux de croissance supérieur à 1,5. Si l'on recherche les marques présentes, les motifs pointent sur des contextes particuliers, par exemple, les phrases avec quantification appartiennent à des sections portant le même trait.

Il est frappant de constater que la mesure virgulot ne fournit pas de bonne discrimination entre classes. En effet, les motifs extraits avec un seuil de fréquence de 500 et un taux de croissance supérieur à 1,5 sont très longs (20 items et plus) et un grand nombre d'items se trouvent partout, ils ne sont donc pas discriminants. Cela revient à dire que le virgulot ne peut pas être caractérisé en soi. Ce résultat corrobore ce qui se lisait « en creux » dans les règles de caractérisation. En effet, un virgulot était jugé correct, non pas parce qu'il portait tel ou tel descripteur, mais par son appartenance à une phrase, un paragraphe et une partie bien marqués. Cela tend à montrer que les modifications les plus importantes concernent les facteurs de lisibilité portés par les hauts niveaux.

Les résultats se lisent facilement sur deux faces : les motifs de marques présentes alternent avec les motifs de marques absentes, caractéristiques soit des segments avec fautes soit des segments sans faute. Les caractérisations interprétables sous conditions apparaissent nettement. Ainsi le descripteur « isopériphérie » est discriminant sur les marqueurs de la sélection section (dont la mesure paragraphe hérite), et non discriminant sur les marqueurs de la sélection paragraphe, à condition que la partie soit cohésive (sans conflit). Autrement dit, si la partie est convenablement marquée, il est indifférent que le paragraphe le soit par isopériphérie.

Les facteurs de cohérence apparaissent également comme décisifs dans les résultats des motifs émergents.

6. Discussion et conclusion

Les conclusions que nous tirons de l'expérience sont provisoires en ce qui concerne l'application, puisque nous avions un jeu de données limité. Concernant la démarche, elle s'avère prometteuse. Concernant les méthodes de fouille, il est intéressant de comparer les motifs caractérisant les deux classes (avec et sans faute) obtenues par les deux méthodes. L'extraction de règles de caractérisation, définissant dans le détail les mesures à fort effectif, s'avère très complémentaire à l'extraction de motifs émergents caractérisant bien les mesures à faible effectif. Les deux méthodes utilisées ont prouvé leur efficacité face à la densité des données fouillées.

La méthode des motifs émergents montre son intérêt pour les mesures à effectifs faibles. Elle fait apparaître clairement des motifs caractéristiques pour la mesure paragraphe, ainsi que pour la mesure partie. Elle est également plus puissante pour faire apparaître des motifs lisibles caractérisant les classes contrastées.

Les expériences menées sont riches d'enseignement, à la fois sur le plan théorique et sur le plan pragmatique. La mesure phrase est la moins modifiée humainement et aussi la moins bien caractérisée automatiquement. La mesure paragraphe est la mieux couverte à travers les deux méthodes d'extraction de connaissances. Aux deux extrêmes, la mesure partie est très informative dans la méthode des motifs émergents, alors qu'elle est inopérante dans la méthode de caractérisation. À l’inverse, la mesure virgulot est intéressante dans la méthode de caractérisation tandis qu'elle est inexploitable par les motifs émergents.

La conclusion la plus nette est que les descripteurs reflétant la cohérence à travers les niveaux sont beaucoup plus puissants que ceux reflétant la cohésion au niveau de chaque mesure. Les caractérisations obtenues mettent en valeur l'importance du contexte ; dans les deux méthodes les données héritées sont capitales. La relation établie entre les descripteurs et le niveau hiérarchique approprié est validée par une bonne convergence des résultats. Reste que nous ne pouvons valider le choix d'un jeu de descripteurs qu'en mettant en concurrence différents jeux de descripteurs plausibles.

Pour le linguiste, la caractérisation des segments sans faute est complétée par la caractérisation des segments avec fautes. Les extracteurs produisent des exemples de conflit local et d'incohérence, ils produisent aussi des séries de descripteurs manquants. On constate que l'incorrection en anglais correspond surtout à une grande sécheresse de style. Les phrases sont mises bout à bout et les parties ne sont pas contrastées. Les réviseurs améliorent la lisibilité des articles en rajoutant des marqueurs ou en déplaçant pour les mettre en meilleure position, notamment des adverbes expressifs. Ils peuvent aussi transformer la voix active en passive et vice-versa. La cohérence est synonyme de compétence dans l'écriture d'articles scientifiques en anglais.

Enfin, nous avons montré par l'utilisation de méthodes de fouille de textes la validité de nos hypothèses sur l'importance de l'organisation qui sous-tend la structure du texte. Aucun marqueur n'est bon ou mauvais en soi, en revanche il est approprié ou non au contexte immédiat et au niveau hiérarchique. Nous avons mis en évidence l'importance du contexte saisi à travers la mise en forme matérielle pour détecter la présence ou l'absence de fautes de compréhension et de style. Ces hypothèses, selon nous fondamentales pour l'objectif visé, ne peuvent pas être prises en compte par les techniques les plus courantes de fouille de textes qui ne tiennent pas compte de l'organisation du document. Ce travail montre que des techniques récentes de fouille de textes peuvent contribuer à apporter des solutions à des tâches aussi difficiles que la détection de fautes de style et illustre la richesse de la complémentarité entre la fouille de textes et la linguistique textuelle.

Les limites de cette expérience concernent la possibilité d'utiliser directement les règles de caractérisation et les motifs émergents pour la prédiction de la classe de nouveaux segments. Même si ceux-ci fournissent de solides bases pour construire des classifieurs (Liu et coll., 2000), la conception de classifieurs fiables reste difficile. C'est une perspective riche, qui constitue un prolongement naturel à ces travaux, dans laquelle nous nous sommes engagés. Concernant le *copyediting*, la caractérisation et la prédiction fiables des segments corrects représenterait un gain de temps appréciable pour l'éditeur. Elle permettrait en effet aux réviseurs de sauter de longs passages de texte bien écrits pour focaliser l'attention sur les segments mal écrits. Une autre perspective concerne le jeu de descripteurs qui peut être raffiné en insistant davantage sur les facteurs de cohésion, notamment pour les mesures supérieures. Nous envisageons d'exploiter les séquences extraites ainsi que les figures de style. D'autre part, il est clair que l'analyse de données plus nombreuses et plus variées sur un nombre suffisamment grand d'articles avant et après révision permettrait de confirmer les conclusions provisoires auxquelles nous avons abouti.

Remerciements

Les auteurs remercient François Rioult, Arnaud Soulet, Leny Turmel et Laëtitia Voisin pour leur contribution à cette recherche.

Bibliographie

Agrawal R., Imielinski T., Swami A. « Mining Association Rules between Sets of Items in Large Databases » in *Proceedings of ACM SIGMOD 93*, pp. 207-216, San Diego, CA: ACM Press, 1993.

Bakhtine M. *Esthétique de la création verbale*. Paris, Gallimard, 1984.

Bastide Y., Taouil R., Pasquier N., Stumme G., Lakhal, L. Pascal « Un algorithme d'extraction des motifs fréquents » *Techniques et Sciences Informatiques*, 21(2), pp. 65-95, 2002.

Bernard P., Dendien J., Lecomte J. « Les ressources de l'ATILF pour l'analyse lexicale et textuelle : TLFi, Frantext et le logiciel Stella » in *6èmes Journées internationales d'Analyse statistique des Données Textuelles (*JADT 2002) 2002.

Biber D. *Variation across Speech and Writing*. Cambridge: Cambridge University Press, 1988.

Boulicaut J. F., Bykowski A., Rigotti C. « Approximation of frequency queries by means of free-sets » in *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD'00, Lyon, LNAI 1910, Springer, pp. 75-85, 2000.

Boulicaut J. F., Bykowski A., Rigotti C. « Approximation of frequency queries by means of free-sets » in *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD'00,* Lyon, LNAI 1910, Springer, pp. 75-85, 2000.

Boulicaut J. F., Bykowski A., Rigotti C. « Free-sets: a condensed representation of boolean data for the approximation of frequency queries » *Data Mining and Knowledge Discovery journal*, 7(1), pp. 5-22, 2003.

Brill E. « Some advances in transformation-based part of speech tagging », *AAAI*, volume 1, pp. 722-727, 1994.

Calders T., Goethals B. « Mining all non-derivable frequent itemsets » in *Proceedings of the Sixth European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD'02*, Helsinki, Springer, pp. 74-85, 2002.

Chali Y., Pascual E., Virbel J. « Text structure Modeling and Language Comprehension processes » in *Proceedings of ALLC-ACH '96*, University of Bergen, 1996.

Chan S., Kao B., Yip C. L., Tang M. *Mining Emerging Substrings* Hong Kong University CSIS technical report TR-2002-11, 25p. 2002.

Charolles, M. « Les études sur la cohérence, la cohésion et la connexité textuelles depuis la fin des années 1960 » *Modèles Linguistiques*, 10 (2) pp. 45-66, 1988.

Cherfi H., Napoli A., Toussaint Y. « Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association », *Actes de la Conférence d'Apprentissage CAp'03* , pp. 61--76, 2003.

Crémilleux B., and Boulicaut J. F. « Simplest Rules Characterizing Classes Generated by δ -FreeSets » in *Proceedings of the twenty-second Annual International Conference of the British Computer Society's Specialist Group on Artificial Intelligence* (ES 02) Cambridge, Springer, pp. 33-46, 2002.

Daille B., Habert B., Jacquemin C., Royauté J. « Empirical Observation of Term Variations and Principles for their Description » *Terminology* 3-2, pp. 197--257, 1996.

De Raedt L., Kramer S. « The Levelwise Version Space Algorithm and its Application to Molecular Fragment Finding » in *Proceedings IJCAI 2001* pp. 853-862, 2001.

Dong G., Li J. « Efficient Mining of Emerging Patterns : Discovering Trends and Differences » *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, ACM Press, pp. 43-52, 1999.

Dong G., Li J., Ramamohanarao K. « Instance-Based Classification by Emerging Patterns » in *Principles of Data Mining and Knowledge Discovery (PKDD'00)*, Cambridge, Springer, pp. 191-200, 2000.

Durand N., Crémilleux B. ECCLAT: a new approach of clusters discovery in categorical data, *proceedings of the twenty-second Annual International Conference Knowledge Based Systems and Applied Artificial Intelligence (ES'02)*, Cambridge, Springer, pp. 177-190, 2002.

Enkvist, N. E. « Coherence, composition, and text linguistics » In Enkvist, N. E. (ed) *Coherence and Composition: A Symposium*, Åbo, Finland, Publications of the Research Institute of the Åbo Akademi Foundation, pp. 11-26, 1985.

Feldman R., Dagan I. « Knowledge discovery from textual databases » In *Proceedings of the International Conference on Knowledge Discovery from Databases*, pp. 16-29, 1995.

Feldman R. et coll. « Text mining at the term level » in *Proceedings of 2nd Conference on Principles and Practice of Knowledge Discovery in Databases PKDD 98*, Nantes, LNAI 1510, Springer, 1998.

Fløttum K. et Rastier F. (eds) *Academic discourse, multidisciplinary approaches*. Oslo, Novus forlag, 2002.

Fontaine L., Kodratoff Y. « Comparaison du rôle de la progression thématique et de la texture conceptuelle chez les scientifiques anglophones et francophones s'exprimant en Anglais » in *Journée de Rédactologie scientifique : L'écriture de la recherche*. Nantes. http://www.lri.fr/~yk, 2003.

Han E-H., Karypis S. G., Kumar V., Mobasher B. « Clustering based on association rule hypergraphs » in *Proceedings of the workshop on Research Issues on Data Mining And Knowledge Discovery, SIGMOD'97*, 1997.

Harris Z. « Discourse analysis » *Language* (28), pp. 1-30, 1952.

Hearst M. « Multi-Paragraph Segmentation of Expository Text » in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, June 1994.

Hyland K. *Hedging in scientific research articles*. Hong Kong, City University of Hong Kong, 1998.

Jacquemin, C., Bourigault, D. « Term Extraction and Automatic Indexing » In R. Mitkov (editor), *Handbook of Computational Linguistics*, pp. 599-615. Oxford, Oxford University Press, 2003.

Kando N. « Text Structure Analysis as a Tool to Make Retrieved Documents Usable » in *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*, Taipei, Taiwan, pp. 126-132, 1999.

Karlgren J. *Stylistic Experiments for Information Retrieval*. PhD Stockholm University, 2000.

Kiyota Y. et Kurohashi S. « Automatic Summarization of Japanese sentences and its Application to WWW KWIC Index » in *Symposium on Applications and Internet (Saint 2001)* San Diego, 2001.

Kodratoff Y. . « Knowledge discovery in texts: a definition and application » in *Proceedings of the International Symposium on Methodologies for Intelligent Systems*, LNAI 1609, pp. 16-29, 1999.

Liu B., Ma Y., Wong C. K. « Improving an Association Rule Based Classifier » in *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 00*, Springer-Verlag, LNAI 1910, pp. 504-509, Lyon, France, 2000.

Lucas N., Nishina K., Suresh K. G. et Akiba T. « *Discourse analysis of scientific textbooks in Japanese : a tool for producing automatic summaries* ». Department of Computer Science Tokyo Institute of Technology, 92TR-0004, 1993.

Lucas N., Crémilleux B., Turmel L. « Signalling well-written academic articles in an English corpus by text-mining techniques » in *Proceedings Corpus Linguistics 2003*, UCREL technical papers 16 Lancaster, pp. 465-474, 2003.

Mannila H., Toivonen, H. Multiple uses of frequent sets and condensed representations, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD 96*, Portland, Oregon, pp. 189-194, 1996.

Mannila H., Toivonen H., « Levelwise search and borders of theories in knowledge discovery » *Data Mining and Knowledge Discovery* 1(3), pp. 241-258, 1997.

Marcu D. *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, Mass., MIT Press, xix, 248 pp. 2000.

Muslea I., Minton S. et Knoblock C. « Active Learning with Strong and Weak Views: A Case Study on Wrapper Induction ». in *Proceedings IJCAI*, 2003.

Parsons, G. *Cohesion and coherence: Scientific texts. A comparative study*. Monographs in Systemic Linguistics, Nottingham, England, Department of English Studies, University of Nottingham, 1990.

Pascual E., Péry-Woodley M-P. « Modèles de texte pour la définition » in *Premières Journées Scientifiques et Techniques du Réseau francophone de l'Ingénierie de la Langue de l'AUPELF-UREF*. AUPELF-UREF, pp. 137-146. 1997.

Pasquier N., Bastide Y., Taouil R., Lakhal L. « Efficient Mining of Association Rules Using Closed Itemset Lattices » *Information Systems*, 24(1) pp. 25-46, 1999.

Péry-Woodley M. P. *Textual designs: signalling coherence in first and second language academic writing*. Notes et documents LIMSI 91-1. Orsay: LIMSI,1991.

Riloff E., Lehnert W. « Information extraction as a basis for high-precision text classification » in *ACM Transactions on Information Systems*, 12(3), pp. 296-333, 1994.

Roche M., Kodratoff Y. « Utilisation de LSA comme première étape pour la classification des termes d'un corpus spécialisé » in *actes de la conférence MAJECSTIC'03*, Marseille, 2003.

Sebastiani F. « Machine learning in automated text categorization » *ACM Computing Surveys*, 34(1), pp. 1-47, 2002.

Soulet A., Crémilleux B., Rioult F. « Condensed Representation of Emerging Patterns » *8th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Lecture Notes in Computer Science*, Sydney, Australia, May 2004.

Salton G. *Automatic text processing: the transformation, analysis and retrieval of information by computer*, Addison-Wesley, 1989.

Swales J. *Genre Analysis : English in academic and research settings*. Cambridge : Cambridge University Press, 1990.

Tierney R. J., Mosenthal J. H. *The Cohesion Concept's Relationship to the Coherence of Text*. Center for the Study of Reading, University of Illinois, Champaign, Ill., 1981.

Turmel L., Lucas N., Crémilleux B. « Extraction d'associations pour la caractérisation de segments de textes en anglais avec et sans faute » in *Actes Cide 6*, Caen novembre 2003.

Voisin, L. Le correcteur automatique d'articles scientifiques en anglais : segmentation et marquage des textes. Rapport de stage sous la dir. de N. Lucas et J. Vergne, GREYC, Université de Caen, 2002.

Zaki M. « Generating non redundant association rules » in *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, ACM SIGMOD'00*, pp. 33-43, 2000.