Détection automatique de la structure organisationnelle de documents à partir de marqueurs visuels et lexicaux

Jean-Philippe Fauconnier, Laurent Sorin, Mouna Kamel, Mustapha Mojahid, Nathalie Aussenac-Gilles

Détection automatique de la structure organisationnelle de documents à partir de marqueurs visuels et lexicaux

Jean-Philippe Fauconnier[1] Laurent Sorin[1] Mouna Kamel[1]
Mustapha Mojahid[1] Nathalie Aussenac-Gilles[1]
(1) IRIT, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex 9
{prénom}.{nom}@irit.fr

Résumé. La compréhension d'un texte s'opère à travers les niveaux d'information visuelle, logique et discursive, et leurs relations d'interdépendance. La majorité des travaux ayant étudié ces relations a été menée dans le cadre de la génération de textes, où les propriétés visuelles sont inférées à partir des éléments logiques et discursifs. Les travaux présentés ici adoptent une démarche inverse en proposant de générer automatiquement la structure organisationnelle du texte (structure logique) à partir de sa forme visuelle. Le principe consiste à (i) labelliser des blocs visuels par apprentissage afin d'obtenir des unités logiques et (ii) relier ces unités par des relations de coordination ou de subordination pour construire un arbre. Pour ces deux tâches, des *Champs Aléatoires Conditionnels* et un *Maximum d'Entropie* sont respectivement utilisés. Après apprentissage, les résultats aboutissent à une exactitude de 80,46% pour la labellisation et 97,23% pour la construction de l'arbre.

Abstract. The process of understanding a document uses both visual, logic and discursive information along with the mutual relationships between those levels. Most approaches studying those relationships were conducted in the frame of text generation, where the text visual properties are infered from logical and discursive elements. We chose in our work to take the opposite path by trying to infer the logical structure of texts using their visual forms. To do so, we (i) assign a logical label to each visual block and (ii) we try to connect those logical units with coordination or subordination relationships, in order to build a logical tree. We used respectively a *Conditional Random Fields* and a *Maximum Entropy* algorithms for those two tasks. After a learning phase, the obtained models give us a 80,46% accuracy for task (i) and a 97,23% accuracy for task (ii).

Mots-clés : discours, structure organisationnelle, mise en forme matérielle, marqueurs métadiscursifs, champs aléatoires conditionnels, maximum d'entropie.

Keywords: discourse, organizational structure, text formating, metadiscursive markers, conditional random fields, maximum entropy.

1 Introduction

La construction automatique de la structure de documents constitue un enjeu majeur en Traitement Automatique du Langage (TAL). En effet, les traitements des modules actuels (e.g : étiquetage morpho-syntaxique, reconnaissance d'entités nommées, etc.) opèrent généralement à un niveau de granularité qui ne prend pas en compte les phénomènes se déroulant au niveau supérieur, tels que les relations entre les sections, titres, paragraphes, etc. (Marcu, 2006). Or, une approche plus globale des textes paraît être une étape nécessaire pour améliorer l'accessibilité des documents (Sorin *et al.*, 2013), la navigation intra-documentaire (Couto *et al.*, 2004), le résumé automatique (Bossard, 2009) ainsi que l'extraction d'information (Fauconnier *et al.*, 2013).

Nous partons du constat qu'un texte peut être segmenté selon trois structures : (i) *la structure visuelle* (segmentation en pages, blocs visuels, etc.), (ii) *la structure logique* (segmentation en titres, paragraphes, etc.), et (iii) *la structure discursive* (segmentation en unités élémentaires et complexes du discours). Les frontières entre ces structures ne sont pas nettement établies dans la littérature. Toutefois, il est admis que ces structures s'échelonnent graduellement dans la compréhension d'un texte et entretiennent des relations complexes d'interdépendance. Par exemple, la mise en forme spatiale d'un texte a des répercussions sur l'interprétation de sa structure logique (Virbel *et al.*, 2005), et une relation logique de coordination entre deux items d'une structure hiérarchique implique une relation rhétorique spécifique (Vergez-Couret *et al.*, 2011).

L'analyse des structures de documents est un sujet traité au sein de la communauté de l'*Analyse de Documents* (conférences ICDAR, IJDAR, etc.). Généralement, cette tâche est vue comme un problème d'analyse syntaxique et un arbre ordonnant les unités du document est attendu en sortie (Mao *et al.*, 2003). Deux domaines sont considérés : l'analyse géométrique (Tokuyasu & Choub, 2001) et l'analyse logique (Klink *et al.*, 2000). Toutefois, les représentations logiques obtenues ne sont souvent pas adaptées à une analyse fine au niveau discursif. Cette difficulté apparaît lorsque des objets textuels complexes conjuguent à la fois mise en forme matérielle et phénomènes discursifs (e.g : structures hiérarchiques imbriquées, définitions, etc.). En outre, les labels logiques ne sont pas toujours fins (Aiello *et al.*, 2002) et il n'existe pas de consensus sur les valeurs qu'ils peuvent prendre. Cela s'explique notamment par un intérêt portant davantage sur l'analyse géométrique (Paaß & Konya, 2012), plus compliquée pour les documents historiques, les lettres, etc., et non sur la construction d'une structure logique en lien avec la structure discursive.

Au sein de la communauté TAL, les dernières années ont montré un intérêt pour les documents (Péry-Woodley & Scott, 2006) et plusieurs approches pour la structuration de ceux-ci en lien avec le discours ont été proposées. Citons la *Document Structure* (Power *et al.*, 2003), le système $DArt_{bio}$ (Bateman *et al.*, 2001) et le *Modèle d'Architecture Textuelle* (Luc & Virbel, 2001). Ces trois approches reposent sur la *Rhetorical Structure Theory* (RST) (Mann & Thompson, 1988). Cependant, bien que ces approches offrent des cadres théoriques poussés, elles ont pour vocation dans leurs implémentations actuelles la génération automatique de textes. L'élaboration de la structure visuelle est généralement faite au travers d'une correspondance entre les structures logiques et discursives données en entrée. À notre connaissance, il n'existe pas d'implémentation opérant le procédé inverse dans une optique discursive.
Notons que d'autres recherches ont visé à produire une structuration des textes dans l'étude de phénomènes locaux, telles que les ruptures thématiques (Choi, 2002; Couto *et al.*, 2004) ou encore les structures fines de texte (Hernandez & Grau, 2005). Cependant, ces approches ne traitent pas la structuration hiérarchique du document dans sa globalité. À l'inverse, (Ratté *et al.*, 2007) proposent un système symbolique pour l'analyse de documents, mais se limitent aux titres, aux chapitres et aux énumérations de premier niveau (non imbriquées) sans proposer de liens entre ces éléments.

Dans ce travail, nous proposons une représentation en arbre du document au travers de relations métadiscursives, appelée structure organisationnelle. Ces relations sont dites métadiscursives car elles ne dépendent pas du contenu propositionnel des unités logiques qu'elles lient (e.g : titres, paragraphes, items, citations, etc.). Nous représentons ces relations par deux relations de dépendance : la subordination et la coordination. L'avantage premier de cette représentation réside dans la détermination du rôle joué par les éléments logiques dans l'ensemble du document. Par exemple, un élément labellisé comme paragraphe peut avoir un rôle d'item dans une structure plus large. Ceci ouvre notamment la voie à une tâche ultérieure visant la reconnaissance de phénomènes complexes agençant plusieurs unités (e.g : structures hiérarchiques à imbrications multiples, etc.). Pour construire cet arbre, notre méthode prend en entrée des documents PDF préalablement traités par une analyse géométrique avec l'outil LA-PDFText (Ramakrishnan *et al.*, 2012) et procède en deux étapes : (i) la reconnaissance des unités logiques au sein des documents et (ii) la construction de l'arbre liant ces unités logiques.

Dans la section 2, nous décrivons les différentes structures et les situons dans les approches existantes. L'arbre en dépendance est décrit en section 3. Nous présentons le corpus en section 4. Les étapes de traitement sont décrites en section 5 et évaluées en section 6. Une discussion est proposée en section 7. Enfin, nous concluons ce travail sur quelques perspectives.

2 Définitions et modèles pour la structuration de documents

Bien que la majorité des travaux s'accorde sur le fait que plusieurs niveaux de structuration existent (visuel, logique et discursif), il n'existe pas de véritable consensus quant aux frontières entre ces niveaux. Nous proposons de définir ces structures et ajoutons la notion de **structure logique profonde** qui correspond à notre structure organisationnelle (Section 2.1). Ensuite, nous montrons dans quelle mesure celle-ci s'intègre dans l'un des modèles préexistants en structuration de documents liés au discours (Section 2.2).

2.1 Définitions des structures

Nous définissons la **structure visuelle** d'un document comme la forme visuelle dans laquelle il apparaît. Les unités visuelles sont identifiées par des indices de nature typographique ou dispositionnelle qui peuvent suivre une convention liée au support, au moyen de production ou au mode divulgation du document. L'*unité élémentaire* est l'alinéa, c'est-à-dire un segment textuel encadré par deux moyens dispositionnels (e.g : retours à la ligne, etc.). Plusieurs alinéas peuvent composer un bloc visuel, dit aussi *unité visuelle complexe*, lorsque l'écart les séparant est plus petit ou égal à l'interligne.

La **structure logique** d'un document se définit comme un niveau abstrait ordonnant le document en *unités logiques élémentaires* et *unités logiques complexes*. Ces unités sont dites logiques, car elles participent à la compréhension du texte en y jouant un rôle métadiscursif, c'est-à-dire indépendant de leur contenu propositionnel. À ce niveau, nous posons pour les besoins de l'analyse deux sous-structures dont la distinction est graduelle :

— La **structure logique de surface** d'un document est composée d'*unités logiques élémentaires*. Ces unités peuvent être un titre, un paragraphe, une note de bas de page, une citation, une référence bibliographique, mais aussi l'alinéa. À ce niveau, le nom de chacune de ces unités dénote le rôle métadiscursif (ou son absence pour l'alinéa) qu'elle joue dans le texte. Cette liste correspond en partie à ce qui est proposé dans les langages de balisage tels que HTML ou LaTeX, où une distinction est faite entre contenu et mise en forme. Pour des raisons pratiques, ces langages permettent de représenter l'alinéa sans pour autant qu'il soit balisé. Notons que (Power *et al.*, 2003) proposent une description des liens entre la *structure logique de surface* et les langages de balisage.

— La **structure logique profonde** correspond à la **structure organisationnelle** d'un document. Celle-ci ordonne les unités logiques élémentaires en *unités logiques complexes* et correspond à l'organisation du document telle que voulue par son auteur. Les unités complexes peuvent être des sections, des structures hiérarchiques, etc., et peuvent s'imbriquer, se chevaucher ou encore se superposer. Au sein de cette structure, un phénomène de changement de rôle peut apparaître. Une unité considérée comme paragraphe lorsqu'elle est prise isolément peut endosser le rôle d'item au sein d'une structure hiérarchique. Ceci survient lorsque ce paragraphe n'est pas mis en forme comme un item et présente des connecteurs tels que « Premièrement », « Deuxièmement », etc. À ce niveau, les unités entretiennent entre elles des relations complexes que nous qualifions de métadiscursives.

Enfin, la **structure discursive** d'un document est la structure qui ordonne son *message*. Les *unités élémentaires* et *complexes* de discours sont liées les unes aux autres par des relations rhétoriques (Mann & Thompson, 1988; Asher, 1993). Il existe une interdépendance forte entre les unités de discours et les unités logiques, car le contexte d'apparition d'une unité influence le rôle qu'elle joue dans la compréhension d'un texte.

2.2 Modèles pour la structuration de documents

Bien qu'initialement orientées pour la génération de textes, les trois approches présentées ci-dessous proposent un cadre théorique utile pour l'analyse de la structuration de documents :

La *Document Structure* de (Power *et al.*, 2003) est définie comme un niveau abstrait et séparé dans la description d'un document. Ce niveau logique se positionne entre la *représentation physique*, qui correspond à la structure visuelle, et le *message*, qui correspond à la structure discursive. Cette théorie voit son origine dans la *Text-grammar* de (Nunberg, 1990) qui différencie la *phrase syntaxique* (contrainte par la grammaire syntaxique) de la *phrase textuelle* (chaîne de caractères commençant par une majuscule et se terminant par un point). La *Document Structure* étend cette distinction et propose plusieurs *unités abstraites* dont l'unité élémentaire est équivalente à l'alinéa. Ces unités sont hiérarchisées selon des critères de composition et forment pour chaque document un arbre en constituants.

Le système de génération automatique de biographies DArt_{bio} proposé par (Bateman *et al.*, 2001) repose sur un modèle distinguant aussi représentation physique (*page layout*), structure logique (*layout structure*) et structure discursive (*rhetorical structure*). La *layout structure* diffère de la *Document Structure* principalement en deux points : (i) le bloc est le seul type d'unité considéré et (ii) l'ordonnancement des blocs ne repose pas sur des critères de composition. Le cadre théorique est riche (e.g : mise en forme avec images, tables, etc.), mais difficilement utilisable hors du contexte des biographies.

Le *Modèle d'Architecture Textuelle* (MAT) (Luc & Virbel, 2001) a pour vocation de représenter les phénomènes architecturaux des textes au travers d'un métalangage Harrissien (Harris, 1971). Ce métalangage permet d'organiser des *objets textuels*, définis comme des segments rendus perceptibles à la surface du texte. Par exemple, la métaphrase suivante :

L'auteur intitule texte(1) par un titre identifié en titre(1)

indique que l'*objet textuel* identifiée comme *titre(1)* endosse le rôle métadiscursif de titre pour *texte(1)*. L'ensemble des métaphrases forment un *graphe architectural* et correspond à la structure logique (de surface et profonde). Le choix des objets textuels et des relations se fait avec la *mise en forme matérielle* qui regroupe les propriétés de réalisation d'un texte.

Nous pensons que le MAT est le modèle le plus apte à représenter la structure organisationnelle. Bien que ce modèle soit nativement orienté vers la génération de textes, il reste un modèle ouvert permettant la description d'*unités logiques complexes* hiérarchiques et transversales. De plus, contrairement à (Power *et al.*, 2003) et (Bateman *et al.*, 2001), une correspondance est faite entre les marqueurs de *mise en forme matérielle* et l'*architecture textuelle* (structure logique).

Ces marqueurs, dits métadiscursifs, se réalisent sous trois formes : (i) les marqueurs dispositionnels tels que les retours à la ligne, les retraits, etc., (ii) les marqueurs typographiques tels que les puces, les numérotations, etc., et (iii) les marqueurs lexicaux correspondant notamment aux marqueurs d'intégration linéaire (MIL) (e.g :« Premièrement », « Deuxièmement », etc.). Ces trois formes peuvent être combinées dans la réalisation d'une même unité logique. Notons qu'une équivalence existe entre la dernière famille et les *introducteurs de cadres* de (Charolles, 1997). Dans la suite, nous utilisons le MAT comme cadre théorique et employons les marqueurs décrits dans ce modèle pour l'analyse de documents.

3 Représentation hiérarchique de la structure organisationnelle

Nous représentons la structure organisationnelle par un arbre en dépendance organisant hiérarchiquement les *unités logiques élémentaires*, telles que les titres, les paragraphes, les items, etc. (Section 2.1). De manière comparable aux travaux de (Choi, 2002) et (Hernandez & Grau, 2005) sur les énoncés, nous proposons de représenter les relations entre unités par des relations de *subordination* et de *coordination*. Une même relation de subordination est partagée par deux unités coordonnées. Et nous posons le nœud factice *texte* comme racine de l'arbre.
Le principe de dépendance suivi consiste à articuler ensemble les unités qui apparaissent liées dans la cohésion du document. Le choix d'une relation entre deux unités se fait sur la base du changement de niveau dans le texte et, parallèlement, par la présence des marqueurs métadiscursifs qui le marquent. Une relation de subordination désigne une descente dans la structure du document, et une relation de coordination lie deux unités partageant le même niveau et le même label.

Cette représentation a l'avantage d'être indépendante de l'ordre *a priori* entre les labels logiques des nœuds. Un élément considéré comme paragraphe peut être subordonné à un item s'il comporte des indices lexicaux qui indiquent cette dépendance. Également, cette représentation ouvre la voie à l'identification ultérieure de phénomènes complexes (e.g : structures hiérarchiques à niveaux multiples, etc.) par parcours d'arbre. Enfin, l'intégration au graphe architectural du MAT sous la forme d'un sous-arbre permettra de typer finement les phénomènes extraits. Toutefois, notons que la représentation proposée dans ce travail est un modèle simplifié et la sémantique réelle des relations n'est pas traitée ici. Citons néanmoins les travaux de (Bouayad-Agha *et al.*, 2000) et (Lüngen *et al.*, 2010) qui abordent cette problématique.

Dans la figure 1, nous proposons deux exemples de correspondance entre un document et sa structure organisationnelle. Dans chacun d'eux, le document est représenté par un schéma où les paragraphes débutent par une majuscule, les items par une puce et les titres sont numérotés. Dans l'arbre, les relations de subordination sont représentées par des arcs continus et les relations de coordination par des arcs en pointillé.

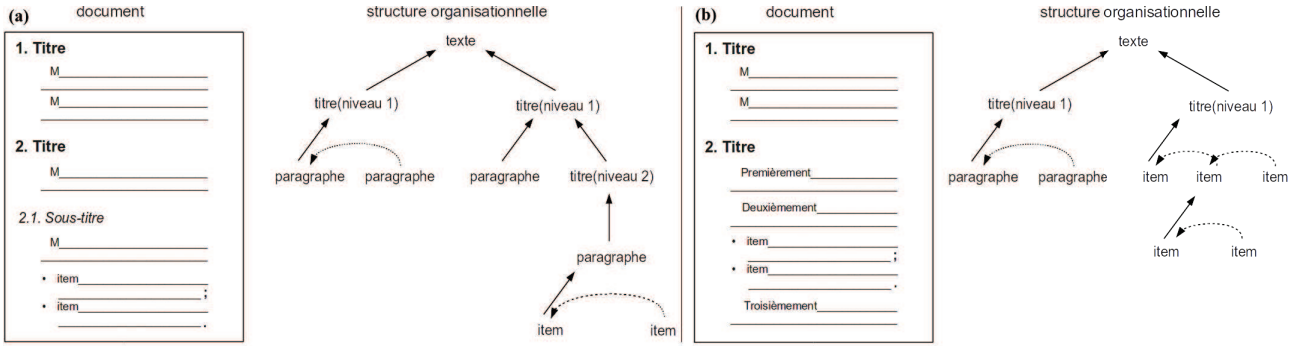

FIGURE 1 – Correspondances entre documents et structures organisationnelles

L'exemple (a) présente une organisation hiérarchique prenant en compte la titraille, ainsi qu'une structure multi-échelle. Une structure multi-échelle, définie par (Ho-Dac *et al.*, 2010), est une *unité logique complexe* qui peut apparaître à tous les niveaux du documents (intra-paragraphique [1], multi-paragraphique, sous-section, section, etc.). Elle se compose d'une amorce, d'une énumération comprenant au-moins deux items, et optionnellement d'une clôture. En organisant leurs items au sein d'une relation d'égalité selon un critère de *coénumérabilité* (implicite ou explicite), les structures multi-échelles participent à la *cohésion textuelle* (Péry-Woodley *et al.*, 2011). Dans (a), la structure multi-échelle est mise en forme visuellement par des indices dispositionnels (les retraits) et des indices typographiques (les puces). L'exemple (b) montre deux structures multi-échelles imbriquées. La première présente des items qui localement endossent le rôle de paragraphe et qui sont introduits par des marqueurs lexicaux. La seconde, qui est imbriquée, est mise en forme comme dans (a).

1. Les structures multi-échelles intra-paragraphiques ne sont pas traitées dans le cadre de ce travail.

4 Construction d'un corpus enrichi visuellement et logiquement

Dans l'objectif d'implémenter des approches par apprentissage supervisé, il a d'abord été nécessaire de construire un corpus riche en marqueurs visuels (typographiques et dispositionnels) et lexicaux. Les corpus LING et GEOP, inclus dans le corpus ANNODIS (Péry-Woodley *et al.*, 2011), ont été choisis comme point de départ car ils présentent deux propriétés : (i) une représentation native au format PDF et (ii) une annotation des structures multi-échelles.

Le corpus LING est constitué de 25 articles scientifiques issus des actes du CMLF 2008 [2]. Le corpus GEOP est constitué de 21 [3] rapports/articles de l'IFRI [4]. Ces deux corpus permettent d'expérimenter deux terrains spécifiques. Au niveau de la structure visuelle, LING présente une mise en forme unifiée (convention du CMLF), alors que GEOP présente des documents très hétérogènes. Au niveau de la structure organisationnelle, LING est relativement complexe, présentant notamment de nombreuses structures multi-échelles imbriquées, tandis que GEOP est plus linéaire.

Nous avons enrichi semi-automatiquement ces corpus par des annotations relatives à (1) leur structure visuelle, (2) leur structure logique de surface et, enfin, (3) leur structure logique profonde. Ce travail a été réalisé en trois étapes successives :

(1) Les documents au format PDF ont été segmentés en blocs visuels en utilisant la segmentation automatique proposée par l'outil LA-PDFText (Ramakrishnan *et al.*, 2012). Cette analyse géométrique repose sur un algorithme qui calcule la proximité entre blocs de mots en prenant en compte leur position mais aussi leurs caractéristiques typographiques locales (fonte, police). Une fois qu'un seuil calculé automatiquement pour chaque page est dépassé, deux blocs de mots sont agrégés. Selon ce principe, cet algorithme segmente de manière ascendante chaque page en une série de blocs visuels.
Toutefois, les blocs proposés par cet outil présentant de nombreuses erreurs (e.g : des paragraphes coupés en deux, inversions dans les blocs, notes de bas de page agglomérées, etc.), un traitement manuel de l'ensemble du corpus a été effectué.
Au terme de cette étape, chacun des blocs visuels contenus dans les documents est caractérisé dispositionnellement et, lorsqu'il s'agit d'un mot, typographiquement. La figure 2 présente un extrait du corpus, où les attributs (x_1, y_1) et (x_2, y_2) représentent respectivement les coordonnées (en pixels) du coin supérieur gauche d'un bloc et de son coin inférieur droit.

```
<page x1="70" y1="71" x2="524" y2="806">
  <chunk x1="70" y1="346" x2="524" y2="360">
    <word x1="106" y1="346"... font="Arial" style="16pt;Bold">Le</Word>
    <word x1="135" y1="346"... font="Arial" style="16pt;Bold">sens</Word>
      ...
  </chunk>
</page>
```

FIGURE 2 – Exemple XML des propriétés visuelles d'un document

(2) Chaque bloc visuel a été annoté avec un label logique élémentaire. Les labels choisis ici sont les titres (de niveau 1 à 3), les paragraphes, les items, les citations, les en-têtes et les pieds de page, les bylines [5], les notes de bas de page et, enfin, les références bibliographiques. Un label *autres* a été choisi pour classer par défaut les blocs non textuels (e.g : images, tables, etc.). Cette étape de classification des blocs visuels comprend deux temps. Premièrement, une annotation a été réalisée avec l'algorithme de similarité textuelle décrit dans (Myers, 1986) en associant les labels logiques du corpus ANNODIS originel aux blocs visuels de notre corpus. Deuxièmement, à l'aide d'une interface en ligne de commande, les labels non traités dans ANNODIS (e.g : en-têtes, notes de bas de page, etc.) ont été ajoutés manuellement.
Au terme de cet enrichissement, des différences significatives (calculées par un χ^2 avec un α à 0,001) apparaissent dans les distributions de labels de LING et GEOP (Tableau 1). Le caractère linguistique de LING implique un plus grand nombre d'items (dont 210 dédiés à l'énumération d'exemples linguistiques tirés de corpus). Son caractère académique implique aussi un grand nombre de citations et de références bibliographiques. Le caractère visuellement hétérogène de GEOP s'observe au travers du grand nombre d'en-têtes et pieds de page, ainsi que dans la présence de nombreuses unités appartenant à la classe *autres* (images et diagrammes). De manière transversale, le paragraphe est l'unité la plus représentée, légèrement en plus grand nombre dans GEOP dont les articles se veulent plus littéraires.

(3) Les deux corpus ont été enrichis par la structure organisationnelle des documents. Cet enrichissement a été effectué en deux temps. Premièrement, des arbres hiérarchiques en dépendance ont été générés à partir d'une grammaire formelle décrivant les relations *a priori* entre les unités logiques élémentaires. Par exemple, un item est subordonné au paragraphe

2. Congrès Mondial de Linguistique Française
3. Sur les 32 articles de GEOP, 21 ont été sélectionnés car considérés comme représentatifs des propriétés visuelles et logiques du corpus.
4. Institut Français de Relations Internationales
5. Le terme byline est un terme générique utilisé pour désigner les lignes de texte en début de document consacrées à l'auteur, sa position et la date.

	h(1,2,3)	para.	item	cit.	en-tête	pied p.	byline	note p.	bibl	autres	Total
LING	304	1241	380	123	45	16	80	394	1173	82	3838
Moy.	12,1	49,6	**15,2**	**4,9**	1,8	0,6	3,2	15,7	**46,9**	3,2	153,5
GEOP	241	1189	72	1	171	257	122	398	25	195	2671
Moy.	11,4	56,6	3,4	0,05	**8,1**	**12,2**	5,8	18,9	1,1	**9,2**	127,1
Total	545	2430	452	124	216	273	202	792	1198	277	6509
Couv.%	8,37	**37,33**	6,94	1,91	3,32	4,19	3,10	12,17	18,41	4,26	100%

TABLE 1 – Distributions des labels logiques au sein de LING et GEOP

qui le précède et deux items contigus sont coordonnés. Deuxièmement, les structures multi-échelles du corpus ANNODIS ont été ajoutées manuellement dans ces arbres. Ainsi, deux paragraphes introduits par des marqueurs d'intégration linéaire peuvent jouer le rôle d'item au sein d'une structure multi-échelle et être subordonnés à un paragraphe apparu précédemment dans le texte endossant le rôle d'amorce.
Le travail de cette troisième étape s'est concentré sur l'étude des relations de dépendance entre les unités logiques élémentaires suivantes : titre, paragraphe, item, citation, référence bibliographique et, enfin, byline. Il nous est apparu que, bien que les en-têtes, les pieds de page et les notes de bas de page participent à la structure organisationnelle des documents, ces unités pouvaient faire l'objet de traitements différenciés, car hors du corps de texte.
Au terme de cette étape, une différence significative apparaît (χ^2 avec α à 0,001) : LING présente des structures organisationnelles beaucoup plus riches avec de nombreuses relations de subordination et de coordination (Tableau 2). Cette différence avec GEOP s'explique encore une fois par le caractère linguistique, pour les subordinations entre paragraphes et exemples linguistiques, et académique, pour les coordinations entre références bibliographiques.

	subordination	coordination	Total
LING	714	2467	3181
Moy.	**28,56**	**98,68**	127,24
GEOP	391	1029	1420
Moy.	18,62	49	67,62
Total	1105	3496	4601
Couv.%	0,24	0,76	100%

TABLE 2 – Distributions des relations de dépendance au sein de LING et GEOP

Notons que, conformément à la licence Creative Commons By-NC-SA 3.0 de ANNODIS, les versions enrichies de LING et GEOP décrites dans cet article sont partagées selon les mêmes conditions [6].

5 Deux tâches pour la détection de la structure organisationnelle

Afin de construire l'arbre correspondant à la *structure organisationnelle* des documents, nous avons décomposé le problème en deux tâches séquentielles :

— **Tâche 1** : labellisation des *blocs visuels* issus de LA-PDFText avec les labels des unités logiques élémentaires (décrites en Section 4) au moyen de marqueurs visuels. Chaque séquence de labels pour un document obtenue en sortie est alors considéré comme la structure logique de surface de ce document.
— **Tâche 2** : construction avec un parseur *shift-reduce* de l'arbre en dépendance reliant les unités logiques élémentaires par les relations *subordination* et *coordination* au moyen de marqueurs visuels et lexicaux, et des labels issus de la Tâche 1. L'arbre en dépendance en sortie est alors considéré comme la structure logique profonde.

Ces deux tâches utilisent respectivement un *Conditional Random Fields* (CRFs), proposé par (Lafferty *et al.*, 2001), et une *régression logistique multinomiale*, introduite en TAL par (Berger *et al.*, 1996) sous le nom de *Maximum d'Entropie* (MaxEnt). Ces deux modèles sont des modèles discriminants, exponentiels et probabilistes qui permettent d'associer à une observation x sa probabilité d'appartenance à un label y, noté $p(y|x)$. Ces deux modèles sont proches, toutefois une différence réside dans le paradigme d'apprentissage ; le CRFs est un modèle graphique et apprend sur des séquences

6. http://github.com/jfaucon/corpus-LING_GEOP

d'observations $X = (x_1, x_2, \ldots, x_t)$ où il peut exister une dépendance statistique entre les labels $Y = (y_1, y_2, \ldots, y_t)$ associés à chaque séquence. Le MaxEnt modélise la probabilité conditionnelle pour une paire (x, y) unique.

Tâche 1. Nous utilisons un CRFs linéaire pour modéliser les dépendances entre les labels logiques y_t et y_{t-1} de deux blocs visuels contigus, ainsi que des informations locales riches sur chaque bloc. La prise en compte des dépendances entre labels est particulièrement adaptée pour capturer et généraliser l'ordre des blocs dans les séquences de documents. Par exemple, un titre est souvent suivi d'un paragraphe, et une référence bibliographique se situe généralement en fin de séquence. Dans sa version du premier ordre, un *champ conditionnel aléatoire* (CRF) prend la forme :

$$p_\theta(y|x) = \frac{1}{Z_\theta(x)} \exp\left(\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right) \tag{1}$$

Tâche 2. Nous utilisons conjointement un parseur *shift-reduce* et un MaxEnt pour modéliser la probabilité conditionnelle d'une paire (x, y) où x est une transition entre deux *unités logiques élémentaires* contiguës et y est l'ensemble {subordination, coordination, $\varnothing$}. Le MaxEnt est un modèle plus adapté aux situations où les distributions sont asymétriques (Malouf, 2002), comme c'est le cas avec la distribution des relations de dépendance. Le MaxEnt prend la forme :

$$p_\theta(y|x) = \frac{1}{Z_\theta(x)} \exp\left(\sum_{k=1}^{K} \theta_k f_k(y, x)\right) \tag{2}$$

Dans les deux modèles, $Z_\theta(x)$ est une constante de normalisation qui assure que la somme des probabilités égale 1, ainsi $Z_\theta(x)$ assure pour le CRF $\sum_y p_\theta f(y_t, y_{t-1}, x) = 1$ et pour le MaxEnt $\sum_y p_\theta f(y, x) = 1$. À chacun des K traits f_k est associé un paramètre θ_k qui donne un poids quant à l'appartenance de x à y.
Théoriquement, le problème dual du MaxEnt, où il s'agit de choisir sous des contraintes calculées à partir des traits la distribution maximisant l'entropie, est semblable à celui du CRFs qui maximise la somme des entropies sous des contraintes calculées identiquement (Ganapathi *et al.*, 2008) [7]. Dans la pratique, l'estimation du vecteur de paramètres θ s'effectue au travers de la maximisation, sans contraintes, de la log-vraisemblance pénalisée sur le corpus d'apprentissage T $(x^{(i)}, y^{(i)})_{i=1}^{N}$. Ainsi, dans les deux modèles, l'estimateur $\hat{\theta}$ est obtenu en (3) par la maximisation de $\mathcal{L}(\theta)$ (4) :

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta) \tag{3}$$

$$\mathcal{L}(\theta) = \frac{1}{N}\left[\sum_{n=1}^{N} \tilde{p}(x^{(i)}, y^{(i)})\ \log\ p_\theta(y^{(i)}|x^{(i)})\right] - \alpha \cdot R(\theta) \tag{4}$$

où $\tilde{p}(x^{(i)}, y^{(i)})$ est la fréquence empirique observée dans T, $R(\theta)$ est un facteur de régularisation et α son coefficient. La propriété de convexité de $\mathcal{L}(\theta)$ assure qu'un *local optimum* est aussi le *global optimum*. Ainsi, une solution unique existe et différents algorithmes itératifs assurent la convergence vers cette dernière. Toutefois, le CRFs nécessite une phase d'inférence à chaque itération pour calculer le gradient de $\mathcal{L}(\theta)$, lié au calcul de la dépendance entre labels. Cette inférence est généralement réalisée avec l'algorithme *forward-backward*, qui présente néanmoins un coût élevé.
Dans notre travail, nous utilisons l'implémentation *crf.sourceforge.net* [8], que nous avons légèrement modifiée pour le support du MaxEnt et la lisibilité des sorties. Pour les deux modèles, nous régularisons $\mathcal{L}(\theta)$ selon la norme L_2 en posant $R(\theta) = \sum_{k=1}^{K} \theta_k^2$. Enfin, nous optimisons $\mathcal{L}(\theta)$ avec l'algorithme LM-BFGS, recommandé pour le MaxEnt par (Malouf, 2002) et pour le CRFs par (Sha & Pereira, 2003).

5.1 Tâche 1 : Labellisation des blocs visuels en unités logiques élémentaires

Pour cette tâche, l'objectif est d'attribuer un label logique aux blocs visuels issus de LA-PDFText sur la base de leurs propriétés de réalisation dans le document. L'hypothèse est double : (i) il est possible d'attribuer un label (e.g : titre, paragraphe, item, etc.) à un bloc visuel à partir de sa mise en forme et (ii) les documents présentent généralement leurs unités logiques selon une séquence générique (e.g : généralement un titre est suivi d'un paragraphe, etc.).

Par conséquent, nous avons défini deux familles de traits : les *traits locaux* qui portent sur les informations locales d'un bloc visuel et les *traits de séquence* qui donnent des informations relatives à la position d'un bloc visuel dans la séquence du document en cours d'apprentissage. Le tableau 3 propose un aperçu synthétique des traits de ces deux familles.

7. Ce problème dual est celui du modèle *Maximum Entropy Markov Models* (HMMs) (McCallum *et al.*, 2000) que le CRFs partage.
8. `http://crf.sourceforge.net`

Les traits locaux se veulent génériques. Ils utilisent des valeurs relatives à chaque document (e.g : une police apparaît majoritairement dans le document, il y a une indentation à gauche, etc.) à la place des valeurs absolues (e.g : une police de taille 10, un retrait de 40 pixels, etc.). Cette manière de procéder permet de se détacher en partie des conventions de mise en forme, qui varient selon le support, le moyen de production ou le mode de divulgation des documents, et d'éviter d'induire des biais dans l'apprentissage. Dans la pratique, ce travail de généralisation des traits nécessite une phase de pré-traitement pour chaque document. Cette phase calcule le mode des variables discrètes (e.g : marges, tailles des polices, etc.) et nominales telles que le style des polices (e.g : Times New Roman, Arial, etc.) ou la présence d'emphase.

Familles	Traits	Informations capturées
Traits locaux	*marges*	Indentation à droite ou à gauche, centrage des blocs, absence d'indentation, etc.
	polices	Présence d'emphases (gras ou italique), taille de la police, etc.
	typographie	Présence de puces, de tirets, de numérotation, d'un « ; » ou « , » en fin de bloc, etc.
	position	Position verticale dans la page (haut, bas) et horizontale (droite, gauche).
	ratios	Ratios de la surface sur la taille de la police, de longueur sur la largeur, etc.
Traits de séquence	*bigrammes*	Considère le label y attribué à l'unité qui précède dans la séquence du document.
	start/end	Présence du bloc en début ou en fin de document.
	contraste	Rupture avec le bloc qui précède (taille, type de police, indentation, etc.).

TABLE 3 – Traits pour la Tâche 1

Les traits *ratios* sont une généralisation statistique de plusieurs caractéristiques locales au sein d'une même formule. Les distributions de valeurs de ces *ratios* sont discrétisées par un découpage en déciles. À chaque décile est associé un trait binaire qui renvoie vrai si le ratio du bloc visuel courant appartient à ce décile.

5.2 Tâche 2 : Construction de l'arbre en dépendance

Pour cette tâche, deux objectifs sont poursuivis : (i) attribuer une relation de dépendance entre deux unités logiques élémentaires et (ii) construire l'arbre correspondant à l'ordonnancement de ces unités au sein de chaque document. Les propriétés hiérarchiques de la représentation de la structure organisationnelle (Section 3) offrent la possibilité de réaliser ces deux objectifs simultanément avec des techniques comparables à celles utilisés en analyse syntaxique.
Nous avons utilisé l'adaptation de l'algorithme *shift-reduce* proposée par (Hernandez & Grau, 2005) pour les énoncés, qui prend en compte les relations de *subordination* et de *coordination*. Le principe de cet algorithme à pile est de parcourir la séquence des unités logiques élémentaires, de gauche à droite, et de chercher le point d'attachement optimal à gauche pour chaque bloc. À chaque étape de la construction de l'arbre, au minimum deux unités sont simultanément inspectées.

Le MaxEnt est entraîné sur trois classes ; la subordination, la coordination et l'absence de relation (notée $\varnothing$). Si une relation de subordination est détectée, l'algorithme descend dans la structure du document (e.g : un paragraphe et un item). Si une relation de coordination est détectée, l'algorithme reste au même niveau dans le document (e.g : deux paragraphes). Enfin, si aucune relation n'est trouvée, l'algorithme remonte dans la structure du document (e.g : un paragraphe et un titre).

Les traits utilisent (i) des informations visuelles (typographiques et dispostionnelles), (ii) des informations lexicales correspondant aux marqueurs d'intégration linéaire, (iii) les labels des unités logiques élémentaires et, enfin, (iv) des informations liés au parallélisme (visuel et lexical) entre unités logiques. Le tableau 4 présente synthétiquement ces traits. Les marqueurs liés aux traits *visuels* sont obtenus de manière similaire à la Tâche 1 (Section 5.1). Les traits *lexique* utilisent une liste prédéfinie de marqueurs d'intégration linéaire. Les traits *labels* et *parallélisme* reposent sur l'hypothèse que nous disposons des résultats produits en sortie de la Tâche 1.

Traits	Informations capturées
visuels	Présence d'indentation, de tirets, de puces, de « : », etc.
lexique	Présence de marqueurs d'intégration linéaire (e.g : *Premièrement*, *Deuxièmement*, etc.).
labels	Paires de labels (e.g : titre-paragraphe, item-item, paragraphe-item, etc.) et égalité de labels
parallélisme	Paragraphe entre deux items visuellement identiques, deux items mais différents , etc.

TABLE 4 – Traits pour la Tâche 2

6 Évaluation

Pour les deux tâches, nous avons procédé à validation croisée (k =10) et présentons les résultats en termes d'exactitude.

Tâche 1. Pour évaluer cette tâche, nous posons la *baseline naïve* consistant à classer tous les blocs visuels en paragraphes, qui forment la classe majoritaire dans LING et GEOP (Section 4). Pour chaque corpus, nous avons effectué l'évaluation selon deux configurations : (i) avec les *traits locaux* seuls (indices typographiques et dispostionnels) et, ensuite, (ii) avec les traits locaux adjoints aux *traits de séquence* (propres au CRFs). Les résultats sont reportés dans le tableau 5.

Approches	LING	GEOP	LING_GEOP
Traits locaux	78,37%	79,97%	73,63%
+ Traits de séquence	**87,18**%	**82,39**%	**80,46**%
Baseline naïve	32,33%	44,51%	37,33%

TABLE 5 – Évaluation pour la labellisation en unités logiques élémentaires (Tâche 1)

Dans la première configuration, les résultats pour LING montrent une difficulté à classer les blocs visuels, avec un léger recul par rapport à GEOP (Δ1,60%). Ce taux bas s'explique notamment par les nombreux exemples linguistiques au sein de LING (Section 4). Ces unités, considérés comme items dans la structure logique de surface présentent des caractéristiques visuelles différentes des items « classiques ». Leur rôle métadiscursif de *citation* implique une indépendance de leur contexte d'apparition. Par exemple, elles ne suivent pas les conventions typographiques des items (e.g : un « ; » en milieu d'énumération et un « . » à la fin) et leur numérotation suit leur ordre d'énonciation dans le document.

Dans la deuxième configuration, la prise en compte de la structure du document permet de palier en partie les variations locales des unités. Cela se traduit par des augmentations significatives (test de Wilcoxon avec α à 0,05) par rapport à la première configuration dans LING (p < 0,01) et GEOP (p = 0,023). Toutefois, pour GEOP, cette amélioration (Δ2,24%) n'est pas aussi élevée que pour LING (Δ8,81%).
Cette différence s'explique notamment par le caractère moins structuré et visuellement hétérogène de GEOP induisant une distribution différente des labels lors de l'apprentissage. Dans les deux corpus, les paragraphes, largement majoritaires, sont correctement classés : F-score de 93,47 pour GEOP et de 90,88 pour LING. Cependant, les nombreuses unités de la classe *autres* (figures, tableaux, etc.) rompent régulièrement la séquence de labels dans GEOP. Ainsi, pour les items, leur nombre restreint et ces variations induisent une diminution (F-score de 26,47 face à 67,58 dans LING). Le même phénomène apparaît également avec les titres de niveau 2 (F-score de 53,17 face à 95,45 dans LING).

Afin de diminuer les variations de distributions dans les corpus d'apprentissage, une évaluation a été faite sur les deux corpus pris conjointement (LING_GEOP). Cette approche avec traits de séquence montre aussi une hausse (Δ6,83%) par rapport aux traits locaux. La figure 3 présente les courbes d'apprentissage avec les traits de séquence sur les trois corpus. Pour obtenir ces courbes, une validation croisée (k=10) à été exécutée pour chaque n de documents choisis aléatoirement dans les 9 ensembles restants. Les résultats semblent indiquer qu'un agrandissement du corpus améliore les scores.

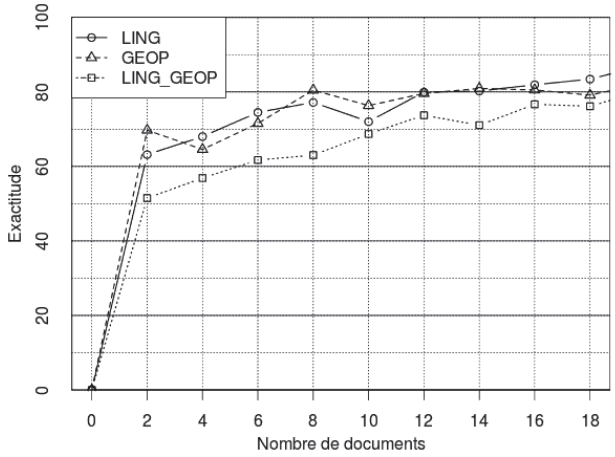

FIGURE 3 – Courbes d'apprentissage pour LING, GEOP et LING_GEOP (Tâche 1)

Tâche 2. Pour l'évaluation de cette tâche, nous proposons une *baseline naïve* consistant à classer aléatoirement les relations de subordination et de coordination liant deux unités logiques. L'approche par *traits* est celle décrite en section 5.2. Nous proposons une comparaison avec une approche par *grammaire formelle* décrivant les règles *a priori* d'organisation d'un document (utilisée pour la construction du corpus en section 4). Les résultats sont reportés dans le tableau 6.

Approches	LING	GEOP	LING_GEOP
Traits	96,41%	**98,45**%	**97,23**%
Grammaire	**96,54**%	98,30%	97,08%
Baseline naïve	40,21%	41,03%	39,79%

TABLE 6 – Évaluation pour la construction de l'arbre en dépendance (Tâche 2)

Dans l'approche par traits, la différence entre LING et GEOP (Δ2,04%) s'explique par la structuration complexe, en termes de dépendances, au sein de LING. Certains documents dans LING montrent des niveaux d'imbrications très profond, notamment par l'utilisation d'exemples linguistiques imbriqués dans des énumérations de définitions. Cela se traduit par des scores différents pour les subordinations (F-score de 91,99 pour LING et de 97,15 pour GEOP), tandis que ceux obtenus pour les coordinations restent relativement équivalents (F-score de 97,15 pour LING et de 98,93 pour GEOP).

Les scores du tableau 6 ne montrent pas de différences significatives entre l'approche par traits et la grammaire (test de Wilcoxon avec α à 0,05). Deux raisons expliquent cela. Premièrement, les relations entre les unités suivent majoritairement les règles définies dans la grammaire. Seuls certains cas (e.g : imbrications profondes, dépendances de longue distance, etc.) permettent de distinguer grammaire et traits. Deuxièmement, cette asymétrie dans la distribution des cas (respectent vs ne respectent pas la grammaire) induit un phénomène d'apprentissage de la grammaire et non des traits considérés comme discriminants (e.g : deux items contigus visuellement différents, un paragraphe indenté, etc.).

Par conséquent, pour mesurer l'apport de l'approche par traits, nous proposons d'évaluer uniquement les cas où la relation entre deux unités diffère de l'ordonnancement *a priori*, c'est-à-dire lorsque la grammaire ne peut fournir la réponse correcte. Les résultats de cette stratégie *traits sur erreurs grammaire* montrent un léger gain qui reste stable au travers des corpus (Tableau 7). Le pendant de cette stratégie consiste à évaluer les traits hors de ces cas. Ces résultats sont ceux de *traits hors erreurs grammaire*, où sont reportés 20 erreurs pour LING, 2 pour GEOP et 12 pour LING_GEOP.

Stratégies	LING	GEOP	LING_GEOP
Traits sur erreurs grammaire	14,54% (16/110)	16,66% (4/24)	14,17% (19/134)
Traits hors erreurs grammaire	99,34% (3051/3071)	99,85% (1394/1396)	99,73% (4455/4467)

TABLE 7 – Deux stratégies pour évaluer les traits face à la grammaire (Tâche 2)

7 Discussion

Les labels logiques utilisés dans les Tâches 1 et 2, ainsi que l'adoption d'une représentation arborée sont des éléments partagés avec l'*Analyse de Documents*. Toutefois, l'objectif de notre travail a été de proposer une représentation en lien avec le discours, mais restant adaptée à l'analyse de documents. Pour cela, il a été choisi de travailler sur le typage des contenus uniquement textuels et une réflexion a été menée sur la différence entre mise en forme et rôle métadiscursif des unités (e.g : un bloc visuel formaté comme paragraphe n'endosse pas toujours le rôle de paragraphe). Les relations de dépendance proposées permettent de représenter cette différence et ouvrent la voie à l'identification de phénomènes discursifs complexes (e.g : énumérations de définitions, etc.). Ces choix, qui ont nécessité l'enrichissement de corpus annotés discursivement, rendent difficile l'utilisation d'outils classiques d'analyse logique. Toutefois, une comparaison externe sur des corpus partageant les mêmes labels textuels est une perspective immédiate à la Tâche 1.

Notre méthode a été testée sur des corpus de natures différentes (une mise en forme unifiée et une structure complexe pour LING, un formatage hétérogène et une structure linéaire pour GEOP). Les résultats obtenus pour les deux tâches sont relativement corrects. Toutefois, des expériences consistant à utiliser en séquence les deux modules ont montré que la Tâche 2 était très sensible au bruit. Ceci constitue pour l'instant un aspect limitatif de notre solution.

Également pour la Tâche 2, il apparaît que les traits lexicaux n'apportent qu'un léger gain par rapport à une approche déterministe. Deux raisons expliquent cette limite. Premièrement, le grain choisi pour l'analyse se limite aux blocs visuellement indépendants et empêche de traiter les cas où les marqueurs d'intégration linéaire sont intra-paragraphiques. Or, ce type de construction est fréquent dans le corpus ANNODIS. Une perspective consistera à travailler avec une granularité plus fine, rapprochant nos travaux de ceux de (Hernandez & Grau, 2005), mais en gardant les marqueurs visuels. Deuxièmement, cette limite s'explique aussi par la variabilité du lexique qui est fonction de la langue et du corpus. Pour améliorer le système, il est nécessaire soit d'étendre les listes données en entrée, ce qui présente un coût, soit d'approcher la tâche de manière plus générique. C'est dans cette dernière direction que nous pensons poursuivre nos recherches. Des traits incorporant des informations syntaxiques pourraient être discriminants. Par exemple, pour l'énumération, il apparaît couramment que la proposition de l'amorce soit liée syntaxiquement aux propositions des items.
Notons que d'autres travaux ont utilisé conjointement mise en forme visuelle et contenu lexical (Klink *et al.*, 2000; Ratté *et al.*, 2007), mais sans proposer une solution traitant les structures imbriquées sur plusieurs niveaux.

8 Conclusion

La contribution de notre approche réside dans la construction automatique de la structure organisationnelle de documents à partir de marqueurs métadiscursifs de nature typographique, dispositionnelle et lexicale. Cette structure est représentée par un arbre en dépendance agençant les unités logiques selon leur label et le rôle métadiscursif qu'elles endossent.
Les perspectives générales de ce travail vont dans deux directions. Premièrement, il est envisagé d'étendre notre approche aux documents numériques tels que les pages HTML de Wikipédia. Ces documents présentent une structuration différente où l'aspect discursif est souvent suppléé par des marqueurs visuels (Bush, 2003). Leur balisage originel permettra de les faire entrer dans le système directement au sein de la Tâche 2. Deuxièmement, l'utilisation de modèles d'apprentissage non-supervisé (clustering) est considérée dans la Tâche 1 afin de ne pas faire d'hypothèse sur les labels logiques. Il s'agira de regrouper les unités de mêmes mises en forme afin d'en prédire le rôle métadiscursif commun dans un second temps.

Références

AIELLO M., MONZ C., TODORAN L. & WORRING M. (2002). Document understanding for a broad class of documents. *International Journal on Document Analysis and Recognition*, **5**(1), 1–16.

ASHER N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Dordrecht.

BATEMAN J., KAMPS T., KLEINZ J. & REICHENBERGER K. (2001). Towards constructive text, diagram, and layout generation for information presentation. *Computational Linguistics*, **27**(3), 409–449.

BERGER A., PIETRA V. & PIETRA S. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, **22**(1), 39–71.

BOSSARD A. (2009). Une approche mixte-statistique et structurelle-pour le résumé automatique. In *Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*.

BOUAYAD-AGHA N., POWER R. & SCOTT D. (2000). Can text structure be incompatible with rhetorical structure ? In *Proceedings of the first international conference on Natural language generation-Volume 14*, p. 194–200 : Association for Computational Linguistics.

BUSH C. (2003). Des déclencheurs des énumérations d'entités nommées sur le web. *Revue québécoise de linguistique*, **32**(2), 47–81.

CHAROLLES M. (1997). L'encadrement du discours. In *Cahier de Recherche Linguistique*, volume 6. Université de Nancy.

CHOI F. Y. Y. (2002). *Content-based Text Navigation*. PhD thesis, the University of Manchester.

COUTO J., FERRET O., GRAU B., HERNANDEZ N., JACKIEWICZ A., MINEL J.-L. & PORHIEL S. (2004). Régal, un système pour la visualisation sélective de documents. *Revue d'intelligence artificielle*, **18**(4), 481–514.

FAUCONNIER J., KAMEL M., ROTHENBURGER B. & AUSSENAC-GILLES N. (2013). Apprentissage supervisé pour l'identification de relations sémantiques au sein de structures énumératives parallèles. In *Actes de la 20e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, p. 132–145.

GANAPATHI V., VICKREY D., DUCHI J. & KOLLER D. (2008). Constrained approximate maximum entropy learning of markov random fields. In *Conference on uncertainty in artificial intelligence (UAI)*.

HARRIS Z. (1971). Structures mathématiques du language. *Dunod. Paris, France*.

HERNANDEZ N. & GRAU B. (2005). Détection automatique de structures fines de texte. In *Actes de la 12e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005)*.

HO-DAC L.-M., PÉRY-WOODLEY M.-P. & TANGUY L. (2010). Anatomie des structures énumératives. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*.

KLINK S., DENGEL A. & KIENINGER T. (2000). Document structure analysis based on layout and textual features. In *Proc. of International Workshop on Document Analysis Systems, DAS2000*, p. 99–111 : Citeseer.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *Department of Computer & Information Science, University of Pennsylvania*.

LUC C. & VIRBEL J. (2001). Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum*, (1), 103–123.

LÜNGEN H., BÄRENFÄNGER M., HILBERT M., LOBIN H. & PUSKÁS C. (2010). Discourse relations and document structure. *Linguistic modeling of information and markup languages*, **1**, 97–123.

MALOUF R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *proceedings of the 6th conference on Natural language learning*, p. 1–7 : Association for Computational Linguistics.

MANN W. & THOMPSON S. (1988). Rhetorical structure theory : Toward a functional theory of text organization. *Text*, **8**(3), 243–281.

MAO S., ROSENFELD A. & KANUNGO T. (2003). Document structure analysis algorithms : a literature survey. In *Electronic Imaging 2003*, p. 197–207 : International Society for Optics and Photonics.

MARCU D. (2006). Automatic discourse parsing. In K. BROWN, Ed., *Encyclopedia of Language and Linguistics*. Elsevier, 2nd edition.

MCCALLUM A., FREITAG D. & PEREIRA F. (2000). Maximum entropy markov models for information extraction and segmentation. In *ICML*, p. 591–598.

MYERS E. W. (1986). Ano (nd) difference algorithm and its variations. *Algorithmica*, **1**(1-4), 251–266.

NUNBERG G. (1990). *The linguistics of punctuation*. Number 18. CSLI Publications.

PAASS G. & KONYA I. (2012). Machine learning for document structure recognition. In A. MEHLER, K.-U. KÜHNBERGER, H. LOBIN, H. LÜNGEN, A. STORRER & A. WITT, Eds., *Modeling, Learning, and Processing of Text Technological Data Structures*, volume 370 of *Studies in Computational Intelligence*, chapter Part V : Document Structure Learning, p. 221–247. Springer.

POWER R., SCOTT D. & BOUAYAD-AGHA N. (2003). Document structure. *Computational Linguistics*, **29**(2), 211–260.

PÉRY-WOODLEY M.-P., AFANTENOS S. D., HO-DAC L.-M. & ASHER N. (2011). La ressource annodis, un corpus enrichi d'annotations discursives. *Traitement Automatique des Langues (TAL)*, **52**(3), 71–101.

PÉRY-WOODLEY M.-P. & SCOTT D. (2006). Computational approaches to discourse and document processing. *TAL*, **47**(2), 7–19.

RAMAKRISHNAN C., PATNIA A., HOVY E. H. & BURNS G. (2012). Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, **7**(1).

RATTÉ S., NJOMGUE W. & MÉNARD P.-A. (2007). Highlighting document's structure. *International Journal of Computer Science & Engineering*, **1**(2).

SHA F. & PEREIRA F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, p. 134–141 : Association for Computational Linguistics.

SORIN L., MOJAHID M., AUSSENAC-GILLES N. & LEMARIÉ J. (2013). Improving the accessibility of digital documents for blind users : contributions of the textual architecture model. In M. A. CONSTANTINE STEPHANIDIS, Ed., *Universal Access in Human-Computer Interaction. Applications and Services for Quality of Life*, p. 399–407. Springer.

TOKUYASU T. A. & CHOUB P. A. (2001). Turbo recognition : a statistical approach to layout analysis. In *Proceedings of SPIE*, volume 4307, p. 124.

VERGEZ-COURET M., BRAS M., PREVOT L., VIEU L., ATTALAH C. *et al.* (2011). The discourse contribution of enumerative structures involving 'pour deux raisons'. In *Procceedings of Constraints in Discourse*.

VIRBEL J., LUC C., SCHMID S., CARRIO L., DOMINGUEZ C., PÉRY-WOODLEY M.-P., JACQUEMIN C., MOJAHID M., BACCINO T. & GARCIADEBANC C. (2005). Approche cognitive de la spatialisation du langage. de la modélisation de structures spatio-linguistiques des textes à l'expérimentation psycholinguistique : le cas d'un objet textuel, l'énumération. In C. THINUS-BLANC & J. BULLIER, Eds., *Agir dans l'Espace*, chapter 12, p. 233–254. Paris : Editions de la MSH.