Anatomie des structures énumératives

Lydia-Mai Ho-Dac[1] Marie-Paule Péry-Woodley[2] Ludovic Tanguy[2]
(1) VALIBEL, UCL et FNRS
(2) CLLE-ERSS, Université de Toulouse
lydia.ho-dac@uclouvain.be, pery@univ-tlse2.fr, tanguy@univ-tlse2.fr

Résumé. Cet article présente les premiers résultats d'une campagne d'annotation de corpus à grande échelle réalisée dans le cadre du projet ANNODIS. Ces résultats concernent la partie descendante du dispositif d'annotation, et plus spécifiquement les structures énumératives. Nous nous intéressons à la structuration énumérative en tant que stratégie de base de mise en texte, apparaissant à différents niveaux de granularité, associée à différentes fonctions discursives, et signalée par des indices divers. Avant l'annotation manuelle, une étape de pré-traitement a permis d'obtenir le marquage systématique de traits associés à la signalisation de l'organisation du discours. Nous décrivons cette étape de marquage automatique, ainsi que la procédure d'annotation. Nous proposons ensuite une première typologie des structures énumératives basée sur la description quantitative des données annotées manuellement, prenant en compte la couverture textuelle, la composition et les types d'indices.

Abstract. This paper presents initial results from a large scale discourse annotation project, the ANNODIS project. These results concern the top-down part of the annotation scheme, and more specifically enumerative structures. We are interested in enumerative structures as a basic text construction strategy, occurring at different levels of granularity, associated with various discourse functions, and signalled by a broad range of cues. Before manual annotation *via* a purpose-built interface, a pre-processing phase produced a systematic mark-up of features associated to the signalling of discourse organisation. We describe this markup phase and the annotation procedure. We then propose a first typology of enumerative structures based on a quantitative description of the manually annotated data, taking into account textual coverage, composition, types of cues.

Mots-clés : Annotation de corpus, organisation du discours, structure énumérative, signalisation.

Keywords: Corpus annotation, discourse organisation, enumerative structure, signalling text structures.

1 ANNODIS : Annotations discursives

1.1 Le projet ANNODIS

Le projet ANNODIS[1] vise la constitution d'un corpus de français écrit enrichi d'annotations concernant le niveau discursif. Tout en se situant dans le sillage de projets d'annotation de relations et structures de discours pour l'anglais – *e.g.* Penn Discourse Treebank (Prasad *et al.*, 2006) – il s'en démarque sur plusieurs points : sa principale originalité est d'aborder l'organisation discursive à partir de deux approches complémentaires – ascendante et descendante ; deuxièmement, les annotations s'appliquent à un

[1]Projet financé pour 3 ans par l'ANR Programme *Sciences Humaines et Sociales*, appel 2007 *Corpus et outils de la recherche en sciences humaines et sociales* : `http ://w3.erss.univ-tlse2.fr/annodis`.

corpus diversifié pour permettre la prise en compte de réalisations discursives variées ; ce corpus fait l'objet de pré-traitements automatiques pour guider les annotations ; enfin, le développement d'outils d'aide à l'annotation et à la navigation constitue un objectif majeur du projet. Cet article expose exclusivement l'approche descendante de ce projet, offrant une première analyse des résultats de la campagne d'annotation des structures discursives de haut niveau[2].

1.2 Annoter des structures discursives multi-échelle

L'approche « descendante » du projet ANNODIS a pour objectif l'identification d'indices de surface signalant une segmentation à différents niveaux de grain. Ces segments, qui vont du paragraphe au texte entier, sont envisagés en relation avec une organisation fonctionnelle : segmentation rhétorique, thématique ou selon différents axes sémantiques (temporalité, spatialité, points de vue, etc.). L'approche que nous développons présente les caractéristiques suivantes :

– elle consiste à aborder le texte d'un point de vue global, à partir de l'hypothèse que la prise en compte des structures de haut niveau a un impact sur l'interprétation locale du texte ;
– elle adopte des méthodes de linguistique de corpus, ce qui implique de se donner les moyens d'analyser un volume conséquent de textes et d'annoter un nombre de structures suffisant pour proposer des généralisations ;
– pour ce faire, elle fait appel à des techniques issues du traitement automatique des langues permettant d'effectuer un pré-marquage automatique des textes, afin de guider l'annotation manuelle des structures ;
– elle s'intéresse à des textes longs non-narratifs, qui ne peuvent se contenter de la seule structuration autour d'un référent principal, et où différents modes de structuration sont susceptibles d'être sollicités et signalés (y compris la structuration dite « logique », cf. « document structure » (Power *et al.*, 2003)).

Les structures discursives recherchées se caractérisant par leur capacité à être perçues avant l'interprétation du contenu propositionnel qu'elles organisent, le rôle de la signalisation à la surface du texte est primordial. De fait, l'existence même des structures recherchées ne se conçoit pas en dehors des indices visuels ou lexicaux qui participent à leur signalisation. Selon cette approche, la perception de l'organisation du discours résulte d'une interaction entre des patrons textuels de haut niveau et le contenu propositionnel.

Prenons par exemple un titre de section annonçant l'évolution de X, suivi de titres de sous-sections localisant les différentes étapes de l'évolution en question : cet ensemble constitue un patron textuel de haut niveau signalant une organisation globale de la section autour de la dimension temporelle, les sous-sections étant en **discontinuité** temporelle les unes par rapport aux autres et en **continuité** du point de vue thématique et du point de vue « textuel » (appartenant à une même section).

Ce type de patron textuel est relativement difficile à repérer dès lors que l'on cherche à décrire des structures indépendantes du découpage visuel. Il peut cependant se réaliser indépendamment de tout alignement sur la structure visuelle du document. Notre premier défi a donc été de définir une structure « annotable » permettant de couvrir les mêmes stratégies organisationnelles que celles observables entre (sous-)sections *i.e.* des structures discursives construites à la fois selon (1) une continuité thématique autour d'un critère X (un référent, une thématique, une entité spatiale-temporelle, un raisonnement, etc.), et (2) des discontinuités internes relatives à un type de trait associable à X (différents sous-thèmes, propriétés, localisations spatio-temporelles, différents points de vue, différentes étapes de construction, de raisonnement, etc.). La *structure énumérative* (ci-après SE) est rapidement apparue comme la structure idéale pour orienter et

[2] Voir Péry-Woodley *et al.* (2009) pour une vue d'ensemble du projet et une présentation de l'approche ascendante qui vise l'annotation des relations de discours entre unités élémentaires de discours.

modéliser l'annotation de ce type d'organisation.

1.3 Les structures énumératives

La structuration énumérative constitue un acte textuel fondamental qui consiste à rassembler des éléments dans un même objet textuel en fonction d'une identité de statut. La structure énumérative consiste ainsi à « transposer textuellement la coénumérabilité des entités recensées par la coénumérabilité des segments linguistiques qui les décrivent, ceux-ci devenant les entités constitutives de l'énumération (les items). L'identité de statut des constituants au sein de l'énumération exprime l'identité de statut des entités recensées dans le monde (cette identité étant, dans les deux cas, la coénumérabilité). » (Luc *et al.*, 2000, p .25).

Une SE se compose nécessairement d'une *énumération* (succession d'items) entourée optionnellement d'une *amorce* et d'une *clôture*. Chaque item a ceci de caractéristique qu'il constitue un segment homogène (*i.e.* une continuité) relativement à un certain critère d'interprétation tout en se définissant par son appartenance à un ensemble de segments entretenant explicitement une relation d'identité de statut. Cette identité relève conjointement des niveaux textuel et propositionnel. Ainsi, les unités co-énumérées apparaissent dans une relation d'égalité vis-à-vis d'un critère X que nous appelons l'**énuméraThème**. Cet énuméraThème correspond à une entité hiérarchiquement supérieure aux items, permettant de rassembler les co-items sous un même « label ». Si l'interprétation de toute SE passe par l'identification d'un énuméraThème, celui-ci n'a pas besoin d'être réalisé explicitement pour exister, pouvant être inféré à partir du contenu des items de l'énumération (Bras *et al.*, 2008, p. 1960).

La figure 1 présente un exemple de plusieurs SE enchâssées dans une SE globale couvrant toute une section de haut niveau. Les vues sont prises sur les lieux de l'annotation : l'interface GLOZZ (voir section 2.2). Sur la gauche figurent le texte de la SE globale et le détail de ses composants ; sur la droite sont détaillées 3 SE dans lesquelles les composants ont été surlignés (items en jaune et amorce/clôture en rose/orange).

Dans cette figure, seules deux SE (surlignées en jaune dans le texte dézoomé) correspondent à l'idée générale des SE, *i.e.* une liste formatée. Le détail de la première SE formatée montre une amorce (*Ces éléments sont*), 5 items pucés et une clôture présentant une deuxième expression de l'énuméraThème (*Ces éléments se sont renforcés...*). La SE la plus à droite de la figure est également une SE complète *i.e.* comportant une amorce (*Par ailleurs, une guerre contre l'Irak pouvait se faire selon trois scénarios.*), trois items, une clôture (*De fait, la première option semblait être la seule...*) et une expression de l'énuméraThème (*trois scénarios*). Ces deux SE, l'une à cheval sur deux paragraphes, l'autre non formatée couvrant un paragraphe entier, montrent bien la variété et la non spécificité des indices pouvant signaler une SE : titres, puces, expressions détachées (adverbiaux circonstanciels *e.g. Depuis quelques années*, ou connecteurs *e.g. Enfin, De fait*), expressions énumérantes (*Le premier*, *La dernière*, etc.).

Comme le montre l'exemple de la figure 1, les SE ne se réduisent pas à une simple question de formatage : la multiplicité de leurs réalisations (listes formatées, découpage en sections, en paragraphes, listes « plates » sans alignement paragraphique) et de leurs rôles fonctionnels (progressions thématiques, segmentation spatio-temporelles, articulations rhétoriques, etc.) en fait un bon point d'entrée dans la complexité de l'organisation discursive. À cette multiplicité s'associe une complexité de signalisation qui ne peut être pensée qu'en termes de faisceaux d'indices diversifiés, à la fois lexicaux, syntaxiques et visuels et non en termes de marqueurs lexicaux dédiés (Péry-Woodley, 2005).

L'annotation des structures énumératives va nous permettre dans un premier temps d'évaluer 4 de leurs propriétés :

1. les textes, surtout de type expositif, font fortement appel à une organisation en SE ;

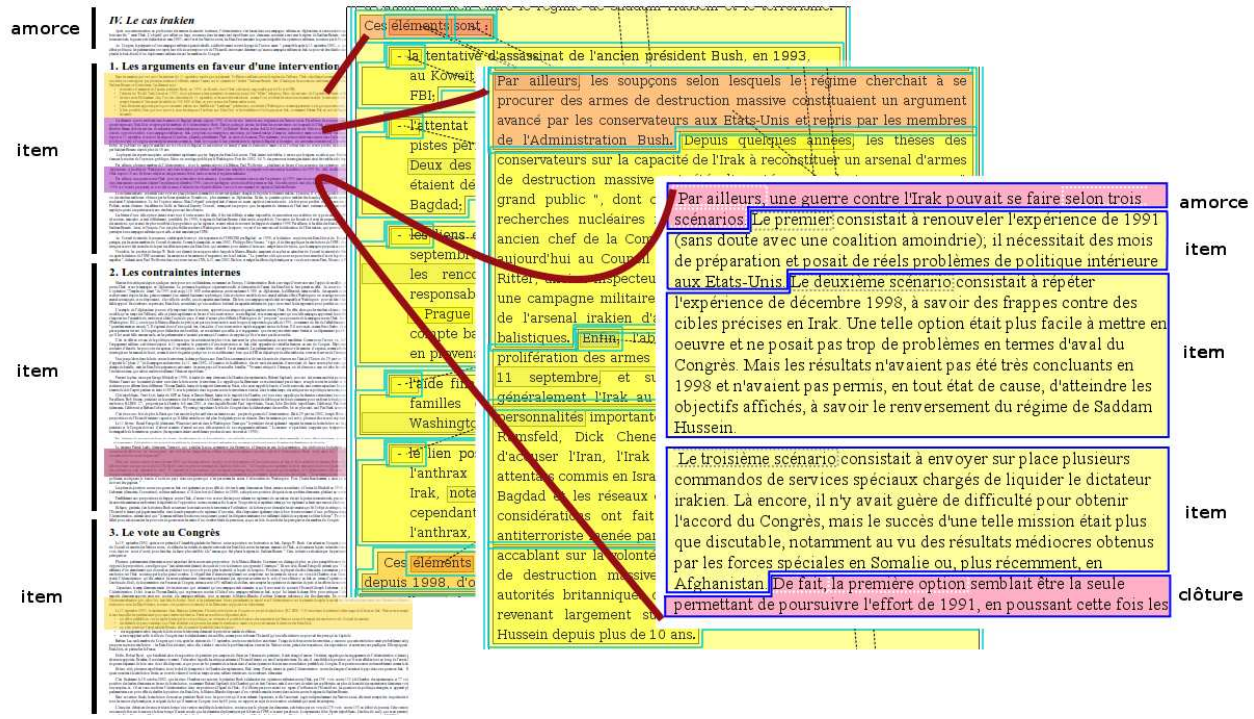

FIG. 1 – Exemple de structures énumératives

2. les SE sont présentes à différents niveaux de grain, du très global (tout un chapitre ou section) au local (quelques propositions) ;
3. ce mode de structuration est associé à une grande variété de patrons textuels : du découpage en sections aux patrons d'amorce et séquences de marqueurs d'items en passant par les listes formatées ;
4. ce mode de structuration est aisément identifiable par le lecteur (Turco & Coltier, 1988, p. 57) et donc l'annotateur.

2 Campagne d'annotation

2.1 Corpus ANNODIS : un corpus diversifié et préparé

Le corpus ANNODIS est un corpus diversifié constitué de textes longs de type expositif présentant une structure visuelle relativement élaborée. Le choix de textes longs de type expositif vise à assurer la présence de structures de haut niveau relativement complexes et signalées à la surface du texte (voir § 1.2). La diversité du corpus nous permet de mesurer la sensibilité des SE à différents usages textuels représentés ici par trois sous-corpus :

– **WIKI** (199 858 mots) : 25 articles longs issus de l'encyclopédie Wikipedia[3] ;
– **GEOPO** (181 414 mots) : 31 articles plutôt polémiques traitant de géopolitique et publiés par l'IFRI[4] ;
– **CMLF** (133 515 mots) : 30 articles scientifiques issus des actes du premier Colloque Mondial de Linguistique Française (CMLF 2008).

[3] `fr.wikipedia.org`.
[4] Institut Français des Relations Internationales, `www.ifri.org`.

Enfin, la présence d'une structure visuelle relativement élaborée et préservée permet la prise en compte d'éléments typodispositionnels jouant un rôle évident dans la signalisation de l'organisation du discours (cf. Power *et al.* (2003)).

La procédure d'annotation manuelle est guidée par un prémarquage automatique de traits qui s'appuie sur la structure de document et sur les sorties d'un étiquetage morpho-syntaxique (TreeTagger) et d'une analyse syntaxique (SYNTEX, Bourigault (2007)). Les éléments prémarqués correspondent à la fois à :

– des expressions associées dans la littérature aux SE : **expressions énumérantes** (séquenceurs tels que *premièrement*, *en second lieu*, *un troisième* X, *d'une part*, etc. décrits par (Turco & Coltier, 1988), (Jackiewicz, 2005) et (Porhiel, 2007) entre autres, auxquels viennent s'ajouter certains connecteurs comme *d'abord*, *ensuite*, *enfin*, *finalement* étudiés par (Bras *et al.*, 2008)) ; **éléments d'amorce** (expressions indéfinies plurielles ou "prospectives" *e.g. selon plusieurs X*, *les X suivants*) **et de clôture** (encapsulations *e.g. ces trois scénarios/de telles propriétés* et connecteurs conclusifs *e.g. En conclusion*)
– des éléments participant de façon plus générale à la signalisation de l'organisation du discours : circonstants détachés en initiale, expressions co-référentielles en position sujet et autres éléments détachés en initiale – *e.g.* tout connecteur, apposition, modalité, etc.
– des patrons ponctuationnels et typo-dispositionnels (listes formatées, patrons de type *[...] : [...] ; [...] et/ou [...]*).

Le prémarquage de ces éléments a un double intérêt : (1) offrir à l'annotateur la possibilité de se dégager d'une lecture linéaire pour aborder le texte avec plus de hauteur, en naviguant entre zones riches en éléments prémarqués ; (2) conduire l'annotateur à prendre conscience de l'existence d'éléments pouvant signaler les structures recherchées (Ho-Dac *et al.*, 2009), éléments qui pourront ensuite être associés au statut d'indice de SE. En plus de ce marquage automatique, la structure visuelle du document est retranscrite dans l'interface d'annotation telle qu'elle apparaît habituellement dans ce type de documents. L'annotation manuelle peut ainsi se baser sur des traits typo-dispositionnels (titres, paragraphes, listes formatées) et lexico-syntaxiques.

2.2 Protocole d'annotation

La campagne d'annotation s'effectue grâce à l'interface GLOZZ créée dans le cadre du projet ANNODIS (Mathet & Widlöcher, 2009). GLOZZ permet de charger des textes enrichis d'annotations ou de prémarquage préalable. Grâce à une « double vue » du texte, l'annotateur peut alterner une approche locale classique et une approche globale qui présente le texte « de haut », lui permettant de visualiser rapidement les différentes mises en forme matérielles, les éléments prémarqués et les annotations réalisées.

La procédure d'annotation est guidée par un *manuel d'annotation* qui définit et illustre chaque type d'unité à annoter et les indices qui peuvent y être associés. L'annotateur, après avoir lu entièrement le manuel d'annotation, charge dans l'interface GLOZZ le texte enrichi, le modèle d'annotation relatif aux SE et les feuilles de styles permettant de colorer les éléments prémarqués. Il se met ensuite à « scanner » le texte à la recherche de zones de concentration d'éléments prémarqués pour, dans un second temps, délimiter précisément les segments par une lecture plus ou moins fine et associer aux SE repérées les indices qui la signalent. Ces indices correspondent soit à un élément prémarqué soit à une nouvelle unité.

L'annotation distingue deux types d'objets : les « unités » (amorce, item, clôture, énuméraThème, indice) et les « schémas » (SE) qui rassemblent plusieurs unités. Les éléments issus de la structure visuelle ainsi que les éléments prémarqués ont également le statut d'unité.

2.3 Accord inter-annotateurs

Trois annotateurs étudiants en sciences du langage (niveau master) ont participé à la campagne d'annotation qui a duré 2 mois, de début septembre 2009 à fin octobre 2009[5]. La présente étude s'appuie sur 65 textes annotés dont 9 avec une annotation multiple. Ces derniers nous ont permis de mesurer l'accord inter-annotateurs sur un total de 361 SE annotées (108+125+128). S'agissant d'une phase de sélection et non de catégorisation nous avons calculé la F-mesure pour comparer les SE identifiées par chacun des annotateurs. Sont considérées comme appariées deux SE couvrant la même zone du texte (avec une variation de quelques caractères au début et à la fin) et possédant les mêmes éléments structurants (amorce, items, clôture[6]). Sur l'ensemble des 9 textes, le taux d'accord moyen est de 0,7. Étant donné la complexité de la tâche, ce score est tout à fait encourageant et montre la stabilité des phénomènes visés par cette phase du projet.

Pour les 56 textes qui n'ont reçu qu'une seule annotation, nous avons mesuré la variation des principales caractéristiques des SE (voir section suivante) entre les annotateurs (*via* des analyses de variance et des tests du χ^2). Aucune de ces caractéristiques n'est significativement liée à l'annotateur, ce qui renforce notre conclusion sur la stabilité de l'objet SE jusque dans ses aspects les plus fins.

3 Analyse des annotations

3.1 Premier inventaire

Si l'on dresse un simple inventaire des annotations effectuées sur les 56 textes qui n'ont reçu qu'une seule annotation, les principales conclusions sont les suivantes :

– sur nos 56 textes, les annotateurs ont identifié un total de 708 SE ;
– tout texte, quel que soit le sous-corpus, contient des SE. Le nombre de SE par texte varie de 3 à 34, avec une moyenne de 12,6 SE par texte ;
– la fréquence relative est de 16 SE pour 10 000 mots ;
– les SEs recouvrent une part importante des textes, avec une moyenne de 46% du texte compris dans au moins une SE[7]. Pour certains textes, ce taux atteint 92%.

Ces premières observations montrent donc que les SE ne sont ni marginales, ni sporadiques, ni réservées à un genre particulier.

3.2 Taille, cardinalité, composition et niveau de grain des SE

Quatre caractéristiques des SE sont distinguées : leur taille (en nombre de mots), leur cardinalité (en nombre d'items), leur composition (présence des éléments optionnels) et leur niveau de grain (interaction avec la structure du document). On va ici examiner leurs variations et tester leur dépendance mutuelle.

Les SE apparaissent tout de suite comme des structures présentant des **tailles** très variées : le nombre de mots que recouvre une SE varie de 8 à 8000 (suivant une distribution de type log-normal) avec une moyenne aux alentours de 400 mots (voir la figure 2).

[5]Ces trois annotateurs pouvaient à tout moment partager leurs impressions sur la tâche d'annotation.

[6]La présence d'un énuméraThème explicite n'a pas été prise en compte pour mesurer l'accord.

[7]Les SE pouvant être enchâssées, certaines zones du texte sont même couvertes par plusieurs SE, mais cet aspect n'a pas été pris en compte dans le calcul.

La **cardinalité des SE** présente, elle, moins de variation : le nombre d'items (seules unités obligatoires) varie de 2 à 48, avec une moyenne de 3,4 items par SE. Malgré 4 *monstres*[8] de plus de 15 items, la majorité des SE comporte entre 2 et 5 items avec une forte proportion d'énumérations à 2 items.

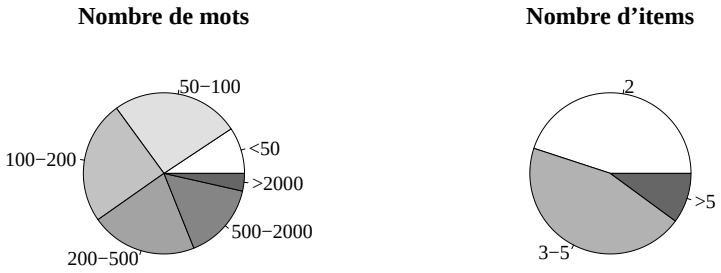

FIG. 2 – Répartition des SE par taille (nombre de mots) et cardinalité (nombre d'items)

Il n'apparaît qu'une corrélation faiblement positive (r = 0,14) entre le nombre d'items et le nombre de mots : ces deux caractéristiques sont relativement indépendantes.

La **composition** d'une SE correspond à la présence des éléments optionnels que sont l'amorce et la clôture, et montre clairement des variations sur l'ensemble de la population : 75% des SE ont une amorce, et 13% une clôture.

Là aussi, aucune variation statistiquement significative n'apparaît entre le nombre de mots et la présence (ou absence) d'une amorce ou d'une clôture, ni entre la composition et le nombre d'items.

Si l'on observe maintenant le **niveau de grain** des SE, mesuré par rapport à leur interaction avec la structure visuelle du document, une grande variété apparaît. Notre première approximation s'appuie sur trois critères :

– le nombre de paragraphes sur lesquels s'étend (au moins en partie) une SE ;
– l'alignement ou non des items avec des titres de section ;
– l'alignement ou non des items avec ceux d'une liste formatée.

Le nombre de paragraphes sur lesquels une SE s'étend varie de 1 à 72, avec une moyenne de 4 paragraphes. Seules 37% des SE se situent à l'intérieur d'un unique paragraphe (grain minimal). A l'inverse, 13% des SE sont alignées sur un découpage du texte en sections (grain maximal), c'est-à-dire que leurs items sont des sections titrées, et plus du quart correspondent à des listes formatées (26,2%).

Des tests statistiques simples (χ^2 et test de Student) permettent de mettre en évidence les liaisons suivantes entre ces niveaux de grain et les autres caractéristiques d'une SE :

– le nombre de mots varie évidemment avec le niveau de grain ($p < 10^{-12}$) : les SE alignées sur les sections sont les plus longues, les listes formatées sont plus courtes que la moyenne, et pour les autres la taille est proportionnelle au nombre de paragraphes ;
– les SE alignées sur les sections ont plus d'items que les autres ($p < 0,01$), plus souvent une amorce ($p < 0,001$), mais moins souvent une clôture ($p < 0,01$) ;
– les SE formatées en listes présentent plus d'items ($p < 0,01$) et plus souvent une amorce ($p < 10^{-12}$) ;
– le nombre de paragraphes est significativement plus important pour les SE sans clôture que pour celles qui en ont une ($p < 10^{-5}$).

Il semble donc que le niveau de grain, qui présente une répartition assez équilibrée des SE de notre corpus, soit la propriété qui possède la plus forte liaison avec les autres caractéristiques (taille, cardinalité, composition). C'est donc sur lui que va se baser notre première typologie des structures énumératives.

[8]La SE de 48 items correspond à une longue liste relatant toutes les étapes du scandale du Watergate, article Wikipedia.

3.3 Typologie des SE

Notre proposition de typologie, basée sur le niveau de grain des SE, est la suivante :

– Type 1 : dont les items correspondent à des sections titrées (93 SE, 13,1%) ;
– Type 2 : dont les items correspondent à des listes formatées (186 SE, 26,3%) ;
– Type 3 : couvrant plus d'un paragraphe sans marques visuelles spécifiques (164 SE, 23,2%) ;
– Type 4 : intra-paragraphiques (265 SE, 37,4%).

Cette typologie établie, nous pouvons maintenant nous intéresser aux fonctionnements des autres caractéristiques des SE : les énuméraThèmes et les indices.

Pour les énuméraThèmes, les SE de type 1 en comportent significativement moins souvent que les autres ($p < 10^{-7}$). En effet, pour ces SE les titres de sections spécifient le critère de coénumérabilité des différents items. À l'inverse, celles de type 2 en comportent plus souvent ($p < 10^{-6}$). Le type 2 apparaît alors comme une sorte d'alternative plus locale au type 1 : le titre de section y étant remplacé par une puce ou un symbole qui n'a pas la capacité d'exprimer le critère de coénumérabilité des items, ce second type requiert davantage l'expression explicite de l'énuméraThème.

Concernant les autres types de SE, il semble que leur spécification se fasse de façon moins affichée (à l'image du saut de paragraphe qui ne signale rien d'autre qu'un acte de segmentation). Il nous faut alors regarder les indices qui ont été associés à ces SE pour trouver une indication quant à leur construction. Pour cela, nous concentrons notre analyse sur les indices trouvés en initiale d'items, considérés comme des alternatives lexico-syntaxiques aux titres et puces des SE de type 1 et 2. Au total, 3 225 indices ont été ainsi associés à une SE : 1 952 éléments prémarqués et 1 273 indices ajoutés par les annotateurs.

Le tableau 1 montre pour chaque type la proportion des items introduits par un indice quel qu'il soit (*i.e.* dont le début correspond avec un indice prémarqué ou ajouté) et la répartition des différents types d'indices associés, en distinguant 6 groupes : les titres de sections (*Titres*), les puces et patrons ponctuationnels (*Ponct.*), les expressions énumérantes (séquenceurs et connecteurs = *S+C*), les adverbiaux en initiale exprimant une circonstance (temporelle, spatiale ou notionnelle = *Circ.*) et enfin les parallélismes syntaxiques(*Parall.*). La catégorie *Autres* recouvre un grand nombre de phénomènes trop isolés pour être pris en compte, ainsi que certains indices non spécifiés par les annotateurs. Le total par ligne peut dépasser 100, certains items étant introduits par plusieurs indices à la fois.

	Nombre d'items			Détails des indices (%)					
	Total	Introduits	%	Titres	Ponct.	S+C	Circ.	Parall.	Autres
Type 1	314	297	94,6	**94,6**	0,7	0,7	0,7	6,1	4,0
Type 2	741	732	98,8	0,0	**98,0**	2,2	1,6	14,9	5,7
Type 3	599	446	74,5	0,0	14,6	**27,8**	**37,2**	15,0	9,9
Type 4	720	478	66,4	0,0	13,2	**46,7**	14,2	20,1	11,9
Total	2374	1953	82,3	14,4	43,4	18,7	12,7	14,8	7,9

TAB. 1 – Signalisation des SE : fréquence des items introduits par un indice et détail des indices

Plus de 80% des items sont introduits par un indice et tous les types d'indices sont représentés, ce qui confirme le rôle de la signalisation dans la définition des SE et la variété des indices impliqués. Si les puces et patrons ponctuationnels constituent le type d'indice prédominant, les autres catégories sont également bien présentes. La quatrième colonne du tableau 1 montre que la progression choisie dans notre typologie scalaire s'associe avec une progression dans la signalisation : pour les types 1 et 2, la signalisation par des

indices visuels est définitoire[9] ; les SE de type 3 et surtout de type 4 font l'objet d'une signalisation moins forte, si l'on prend en compte tous les types d'indices.

Si l'on regarde maintenant le détail des indices, différentes répartitions apparaissent : les expressions énumérantes (S+C) constituent le type d'indice le plus utilisé pour introduire les SE de type 4 ; elles s'associent aux circonstants pour introduire les SE de type 3. Le parallélisme syntaxique est davantage présent dans les SE de type 4.

Quatre conclusions découlent de ces résultats :

– le niveau de grain des SE (traduit par son type) implique des stratégies différentes pour leur insertion dans le texte, avec alternance des marques visuelles (titres, listes) et des marques lexico-syntaxiques ;
– les expressions énumérantes, qui indiquent essentiellement le statut d'item plus que le critère de coénumérabilité, apparaissent à tous les niveaux ;
– les circonstants, qui spécifient le critère de coénumérabilité sans en indiquer explicitement le statut, s'associent davantage à des SE de haut niveau, multiparagraphiques ;
– les parallélismes signalent de façon beaucoup plus locale le critère de coénumérabilité et le statut d'item, sans alignement sur la structure visuelle du document.

Il est important de noter que les éléments prémarqués utilisés lors de cette campagne ont pour vocation d'être affinés. Ces premiers résultats nous permettent déjà d'envisager un prémarquage plus précis des textes, voire de viser un repérage automatisé de ces structures dans la prochaine étape de notre travail. Ils nous encouragent également à examiner d'encore plus près ces structures textuelles. La typologie proposée, comme on le voit, n'est qu'une première façon d'identifier des variations pertinentes dans leur fonctionnement. Une deuxième étape consistera à dresser une typologie fonctionnelle des SE (SE rhétoriques, thématiques, temporelles, etc.). Le croisement de ces deux typologies nous permettra de corréler configurations d'indices, fonctions discursives et niveau de grain.

4 Conclusion

Les premiers résultats de notre projet d'annotation manuelle systématique des structures énumératives dans un corpus diversifié nous ont permis de proposer une typologie préliminaire basée sur l'analyse quantitative des structures et de leurs constituants. Au-delà de cette typologie, nous pouvons à ce stade porter un regard sur l'expérience d'annotation dans son ensemble. Le choix des structures énumératives comme point d'entrée dans la complexité de l'organisation des textes s'avère fructueux : en tant que structures fondamentalement multi-échelle, elles constituent un objet privilégié pour l'étude de l'organisation textuelle à différents niveaux de grain ; en tant qu'objets textuels facilement identifiables, elles permettent de réaliser une annotation fiable et stable ; enfin, la couverture textuelle de ces structures montre que loin d'être un phénomène marginal, elles représentent bien une stratégie de base pour la mise en texte. La mise en place d'une méthodologie et d'un outillage permettant la prise en compte des aspects visuels et de la structuration « logique » des documents, indispensable pour ces structures, est un pas en avant pour l'analyse et l'exploitation de l'organisation textuelle d'une façon générale. Signalons également, toujours sur le plan des méthodes, la mise en œuvre de techniques de traitement automatique pour le pré-marquage, ainsi que les possibilités de visualisation de ce pré-marquage grâce à l'interface GLOZZ. Les indices annotés vont maintenant faire l'objet d'analyses statistiques pour mettre en évidence les configurations récurrentes qui sont les marqueurs de SE à proprement parler. Ces configurations seront ensuite évaluées

[9] Seuls les indices validés manuellement sont comptés : certains ont échappé aux annotateurs, ce qui explique les pourcentages d'indices inférieurs à 100%. Le typage est par contre basé sur l'exploitation systématique de la structure du document.

en tant que marqueurs de segmentation pour repérer des structures hiérarchiques. Par ailleurs, d'autres configurations impliquant d'autres éléments (temps verbaux, expressions (co-)référentiels, constructions spéciales, longueur des paragraphes, etc.) seront définies permettant de délimiter différents items et par extrapolation différents segments de discours. Enfin, nous espérons corréler ces investigations à d'autres études linguistiques grâce à la mise à disposition prochaine du corpus annoté et des outils et ressources construits au cours du projet.

Remerciements

Les auteurs tiennent à remercier les autres membres du projet ANNODIS, et plus spécifiquement : Cécile Fabre et Josette Rebeyrolle pour leur participation centrale à ces travaux, Nikola Tulechki pour le prétraitement des corpus, sans oublier nos fidèles annotateurs de SE : Maud Colleter, Guillaume Carbou et François Morlane-Hondère.

Références

BOURIGAULT D. (2007). Un analyseur syntaxique opérationnel : Syntex. Mémoire d'HDR, Université de Toulouse.

BRAS M., PRÉVOT L. & VERGEZ-COURET M. (2008). Quelle(s) relation(s) de discours pour les structures énumératives ? In *Actes du Congrès Mondial de Linguistique Française*, p. 1945–1964, Paris.

HO-DAC L.-M., FABRE C., PÉRY-WOODLEY M.-P. & REBEYROLLE J. (2009). Corpus annotation of macro-discourse structures. In *Proceeding of CILC09*, Murcia, Spain.

JACKIEWICZ A. (2005). Les séries linéaires dans le discours. *Langue française*, **148**, 95–110.

LUC C., MOJAHID M., PÉRY-WOODLEY M.-P. & VIRBEL J. (2000). Les énumérations : structures visuelles, syntaxiques et rhétoriques. In *Actes de CIDE 2000 (Colloque International sur le Document Électronique)*, p. 21–40.

MATHET Y. & WIDLÖCHER A. (2009). La plate–forme GLOZZ : environnement d'annotation et d'exploration de corpus. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

PÉRY-WOODLEY M.-P. (2005). Discours, corpus, traitements automatiques. In A. CONDAMINES, Ed., *Sémantique et corpus*, p. 177–210. Paris : Hermès.

PÉRY-WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO-DAC L.-M., DRAOULEC A. L., MATHET Y., MULLER P., PRÉVOT L., REBEYROLLE J., TANGUY L., VERGEZ-COURET M., VIEU L. & WIDLÖCHER A. (2009). Annodis : une approche outillée de l'annotation de structures discursives. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

PORHIEL S. (2007). Les structures énumératives à deux temps. *Revue Romane*, **1**(42), 103–135.

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

PRASAD A., MILTSAKAKI R., DINESH E., LEE N., JOSHI A. & WEBBER B. (2006). *Penn Discourse TreeBank 1.0 Annotation Manual*. `www.seas.upenn.edu/~pdtb/`.

TURCO G. & COLTIER D. (1988). Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire. *Pratiques*, **57**, 57–79.