Les chaînes topicales dans la ressource ANNODIS

Silvia Federzoni, Lydia-Mai Ho-Dac, Josette Rebeyrolle

Les chaînes topicales dans la ressource ANNODIS

Silvia Federzoni[1], *Lydia-Mai* Ho-Dac[1], et *Josette* Rebeyrolle[1]

{silvia.federzoni,lydia-mai.ho-dac,josette.rebeyrolle}@univ-tlse2.fr

[1]CLLE-ERSS (UMR 5263, CNRS & Université de Toulouse Jean Jaurès), Maison de la Recherche, 5 Allée Antonio Machado, 31058 Toulouse cedex 9, France

Résumé. Le modèle d'annotation en structures multi-échelles de la ressource ANNODIS est centré sur deux stratégies discursives et deux structures susceptibles d'apparaître à de très hauts niveaux d'organisation : l'empaquetage, réalisé par les structures énumératives ; le chaînage, réalisé par les chaînes dites topicales. Alors que les structures énumératives ont fait l'objet de nombreuses études, l'analyse des « chaînes topicales » restait à faire. C'est à cette analyse qu'est consacrée la présente étude. Cette étude se donne pour objectif de comparer les chaînes topicales, telles qu'elles ont été annotées dans la ressource, avec les chaînes de référence, telles qu'elles sont décrites dans la littérature. Les résultats de la comparaison montrent la pertinence et l'intérêt de la méthode d'annotation utilisée pour la détection de segments textuels organisés autour d'un même référent.

Abstract. Topical chains in the ANNODIS resource. The ANNODIS annotation model focuses on two discursive strategies and two structures likely to appear at a very high level of organisation: packaging, carried out by the enumerative structures; chaining, carried out by topical chains. While enumerative structures have been the subject of numerous studies, the analysis of topical chains remains undone. The objective of the present study is to do it by comparing topical chains, as annotated in the ANNODIS resource, with reference chains, as described in the literature. The results of such a comparison show the relevance and interest of the ANNODIS annotation method.

1 Introduction

Le présent article a pour objectif de décrire les résultats d'une première exploitation de structures discursives annotées dans la ressource ANNODIS et présentées, par les concepteurs de cette ressource, comme des « chaînes topicales » (Péry et al. 2011, Asher et al., 2017). Il s'agit plus précisément de segments mettant en jeu la notion de *topique* et s'appuyant sur les notions d'*à propos* et de structure informationnelle (Lambrecht 1994). Dans le cadre du projet ANNODIS, annoter les « chaînes topicales » d'un texte a consisté à délimiter des segments reconnaissables sur la base du fait que les propositions qui les composent ont en commun un même topique. Le terme « reconnaissable » est essentiel dans le cadre de ce projet où l'annotation a été réalisée par des annotateurs non-experts en

suivant une méthode dite « descendante » dont l'objectif était de « sonder les intuitions des annotateurs sur l'existence de segments définissables en termes de référent topical » (Ho-Dac et Péry-Woodley, 2014). Un autre objectif était de tester une caractéristique essentielle de ces structures qui nécessitent d'être reconnues par le lecteur pour qu'il y ait compréhension du texte – caractéristique que l'on trouve très régulièrement associée à ces structures, comme le dit Webber (in Weber et al., 2012 : 439) : « Discourse structures are the patterns that one sees in multi-sentence (multi-clausal) texts. Recognizing these pattern(s) in terms of the elements that compose them is essential to correctly deriving and interpreting information in the text. » En raison de la place accordée, dans le projet ANNODIS, à la signalisation de l'organisation du discours, les chaînes topicales (désormais CT) ont été systématiquement associées, par les annotateurs, aux indices de surface qui leur ont permis de les reconnaître. Dans cette étude, nous proposons une première analyse de ce jeu d'annotations en les confrontant aux descriptions proposées dans les travaux sur les expressions coréférentielles et les chaînes de référence (désormais CR). Cette confrontation présente l'intérêt de fournir des repères nécessaires pour mettre de l'ordre dans la variété des annotations produites dans la ressource ANNODIS, et ce dans un contexte où les travaux portant sur les chaînes de référence en français sont particulièrement dynamiques (comme le montrent les numéros de revue récents consacrés à ces questions : numéro 195 de *Langue Française* et numéro 195 de *Langages*, ainsi que la récente mise à disposition de la ressource DEMOCRAT (Landragin, 2015)).

Après une description détaillée des spécificités de la ressource ANNODIS concernant l'annotation des chaînes topicales (cf. partie 2), la partie 3 se donne pour objectif d'exposer non seulement la méthode utilisée, mais également les critères de comparaison des chaînes topicales, telles qu'elles ont été annotées dans la ressource, avec les chaînes de référence, telles qu'elles sont décrites dans la littérature. La dernière partie est consacrée à la présentation des principaux résultats de la comparaison.

2 Les chaînes topicales de la ressource ANNODIS

La ressource ANNODIS, constituée dans le cadre du projet ANR ANNODIS (ANNOtation DIScursive), est un corpus français enrichi de deux types d'annotation manuelle, ayant abouti à la constitution de deux ressources distinctes : une ressource annotée en relations rhétoriques et une ressource enrichie de l'annotation de structures multi-échelles, les chaînes topicales (CT) et les structures énumératives (SE). La présente étude repose sur la ressource annotée en structures multi-échelles et plus précisément sur les annotations en CT, qui jusqu'à présent n'avaient pas été décrites. Le segment de texte grisé en (1)[1], ci-dessous, fournit l'exemple d'une CT, dont le topique est *acteurs locaux dans la résistance de petits pays face à la domination des États-Unis* :

(1) En réalité, plus les États-Unis s'approcheront du territoire tenu par l'ennemi, plus celui-ci se montrera efficace, sous l'effet de facteurs politiques, physiques et technologiques combinés. [...] Les facteurs essentiels sont ici les suivants. En premier lieu, la guerre a en général pour les acteurs locaux un intérêt politique de premier ordre, souvent bien plus important que celui des États-Unis. Leur tolérance à la souffrance est donc plus grande. En deuxième lieu, en dépit de leur taille réduite, ces acteurs supplantent d'ordinaire les États-Unis dans une ressource précise : le nombre d'hommes en âge de combattre. [...]. Troisièmement, les « locaux » disposent en général d'un avantage : ils jouent à domicile. [...]. Quatrièmement, nombre des chefs militaires de ces États ou entités ont été formés dans le monde développé [...] Cinquièmement, l'arsenal nécessaire [...]. En outre, la diffusion des capacités économiques et technologiques civiles [...]. Tous ces facteurs se renforcent et contribuent à créer une "zone contestée". [geop_9CT_coder2_1326102481701]

Les textes qui composent la ressource ANNODIS sont des textes longs, de type non narratif relevant de trois genres textuels : des rapports et d'articles de géopolitique publiés par l'Institut Français des Relations Internationales (IFRI), d'où est tiré l'exemple (1), des articles de recherche en linguistique, issus du Congrès Mondial de linguistique Française (CMLF 2008) et des articles encyclopédiques complets tirés de Wikipédia. Ces textes sont organisés en trois sous-corpus auxquels nous renverrons au moyen des étiquettes suivantes : GEOP, LING et WIK2 (Péry-Woodley et al., 2011).

Au moment d'exploiter la ressource ANNODIS pour étudier les chaînes topicales, il est nécessaire de tenir compte des spécificités de cette ressource : spécificité d'abord du modèle et de la méthode d'annotation, spécificité, ensuite, de la notion de chaîne, et enfin spécificité du niveau d'expertise des annotateurs. Avant de fournir un premier aperçu des données à disposition, les sections qui suivent seront consacrées à la présentation détaillée de ces spécificités.

2.1 Modèle et méthode d'annotation des CT

Comme nous l'avons dit en introduction, un des enjeux du projet ANNODIS était de tester le caractère « reconnaissable » de certaines structures discursives, avec l'hypothèse que cette détection « de haut » se ferait en parallèle d'un processus de construction du sens par compositionnalité défendu par les théories du discours qui définissent les structures comme des ensembles d'unités élémentaires reliées par des relations de cohérence e.g. les modèles de la RST (Mann & Thompson, 1988 ; Carlson et al., 2003), du Penn Discourse Treebank (Prasad et al., 2006) et de la SDRT (Asher & Lascarides, 2003).

Pour tester le caractère « reconnaissable de haut », une méthode dite « descendante » a été développée pour l'annotation des structures multi-échelles considérées dans le projet, à savoir les CT et les structures énumératives. Contrairement aux méthodes dites « ascendantes » qui consistent à détecter les unités minimales — par exemple les unités élémentaires de discours ou les expressions référentielles — pour les relier dans un second temps au sein d'une même structure — par exemple une séquence narrative ou une CR, l'annotation « descendante » consiste à détecter la structure avant d'identifier ses composants. Ainsi, pour les CT, l'annotateur avaient reçu la consigne suivante (Colléter et al., 2012) :

« Définition :

Une CT est un segment qui se caractérise par le fait que la majorité des propositions qui le composent ont pour objet (parlent de, sont à propos de, apportent des informations au sujet de) un seul et même référent. Selon cette définition, l'expression de ce référent commun doit passer nécessairement par le sujet grammatical.

Veuillez noter tout de même qu'une CT n'est pas nécessairement composée uniquement de propositions portant sur le référent qui fait l'unité du segment. En effet, des commentaires ou illustrations, par exemple, peuvent être insérés à l'intérieur d'une CT.

Annoter une CT :

Annoter une CT consiste à identifier ce qui fait son unité référentielle ainsi que les indices vous ayant permis de le repérer. »

Nous nous arrêterons sur la dernière phrase de la consigne qui demande à l'annotateur d'indiquer les marques de surface (appelés « indices ») ayant participé à l'identification, à la « reconnaissance » d'une structure. C'est sur ces indices que se fonde l'analyse du rôle de la signalisation, à la surface du texte, dans l'interprétation du contenu propositionnel des structures discursives. Ainsi, dans l'exemple (1) reproduit partiellement ici sous (1'), les cinq indices soulignés en gras ont été identifiés par l'annotateur comme signalant la CT concernant « les acteurs locaux » :

(1') En premier lieu, la guerre a en général pour **les acteurs locaux** un intérêt politique [...]. **Leur tolérance** à la souffrance est donc plus grande. En deuxième lieu, en dépit de **leur**

taille réduite, **ces acteurs** supplantent [...]. Troisièmement, **les "locaux"** disposent en général d'un avantage : **ils** jouent à domicile. [...], **les acteurs locaux** ont fait un travail similaire sur leur propre pays. **Ils** connaissent [...].

Ce type d'annotation « d'en haut » a pu être mis en œuvre grâce à la plateforme d'annotation GLOZZ (Mathet et al., 2009). Cet outil d'annotation a l'avantage de fournir une « double vue » du texte : une visualisation « de haut » et une visualisation « locale » lisible et annotable. L'outil est particulièrement adapté à ce type d'annotation parce qu'il est conçu pour permettre l'annotation de textes longs en conservant le plus fidèlement possible leur mise en forme caractérisée en particulier par la présence de divers niveaux de titres, de listes, de saut de paragraphes, mais aussi d'italiques, de gras. Une autre caractéristique de cet outil est sa capacité à traiter des textes préalablement annotés ou prémarqués. Cette caractéristique a également été exploitée dans le cadre du projet ANNODIS puisque les textes fournis aux annotateurs avaient fait l'objet d'un prémarquage automatique, consistant à mettre en évidence un certain nombre d'indices considérés comme des marques discursives participant à la signalisation du discours. Parmi ces éléments prémarqués, certains contribuent au signalement de la continuité référentielle, comme par exemple les formes pronominales ou possessives en position sujet (e.g. « ils » et « Leur tolérance » en (1')), les SN démonstratifs en position sujet (« ces acteurs » en (1')) et les SN sujet dont la tête reprend un nom déjà introduit dans la section en cours ou le titre de section (la 2^{e} mention de l'expression référentielle définie plurielle « les acteurs locaux » en (1')).

Associé aux possibilités de visualisation offertes par l'outil Glozz, le prémarquage rend possible le parcours des textes « de haut » et facilite l'identification de zones à forte concentration d'éléments prémarqués. Une fois la zone identifiée, l'annotateur peut effectuer un « zoom » pour accéder à la zone concernée grâce à la visualisation locale et procéder à l'annotation proprement dite, en délimitant le segment correspondant à la CT et en indiquant les indices ayant conduit au repérage de la structure. Ces possibilités permettent à l'annotateur de s'affranchir d'une lecture complète et linéaire du texte, qui est en revanche indispensable pour l'annotation des CR.

La méthode généralement adoptée pour l'annotation des CR suit en effet une approche ascendante. Cette approche comprend deux étapes : la première consiste à identifier les expressions référentielles susceptibles de constituer le maillon d'une CR et la seconde à mettre en relation ces expressions – cette mise en relation peut se faire soit en associant les expressions identifiées à une structure, soit en leur attribuant un identifiant commun. Ce type d'annotation exige une lecture précise de l'ensemble du texte à annoter. L'annotation descendante, en revanche, en ce qu'elle permet un survol général du texte orienté par la recherche de zones présentant une forte présence d'éléments prémarqués, n'exige pas une lecture exhaustive du texte à annoter. De ce fait, elle est moins chronophage. Mais est-elle aussi efficace pour mettre au jour les expressions référentielles qui s'inscrivent dans des chaînes ? C'est la question à laquelle nous allons essayer de répondre. Dans cet article, l'enjeu est en effet précisément d'évaluer dans quelle mesure une telle méthode entraîne ou non une dégradation de la qualité des annotations qu'elle permet d'obtenir.

2.2 Un définition de chaîne adaptée à des annotateurs non experts

Le modèle d'annotation défini pour l'annotation des chaînes topicales se distingue de ceux utilisés dans beaucoup de campagnes d'annotations portant sur l'organisation du discours sur deux points : d'abord, il s'agit d'un modèle moins complexe, mais surtout et corollairement, il permet de confier l'annotation à des annotateurs non experts. Rappelons que la tâche des annotateurs consiste ici, une fois la structure identifiée et délimitée, à indiquer les éléments qui signalent une continuité, en les catégorisant comme des indices. Concernant la délimitation de la structure, aucune consigne relative à l'ouverture et la

fermeture des CT n'a été fournie. Les seules indications données aux annotateurs sont qu'une CT doit être constituée d'au moins deux phrases et ne peut pas dépasser une section. Concernant l'indication des indices, aucun nombre minimal d'indice à identifier n'était fixé et il était demandé aux annotateurs de se focaliser sur l'identification de ceux-ci sans renseigner leur nature de façon précise comme c'est le cas dans la plupart des campagnes d'annotations où les annotateurs sont des experts. La possibilité d'indiquer, pour chaque indice, sa nature était toutefois offerte, mais sans suivre un modèle théorique prédéfini si ce n'est celui esquissé par le prémarquage. Pour illustrer, nous nous appuierons sur l'exemple (2) extrait de l'article Wikipédia intitulé « Opération Barbarossa » dans lequel la catégorie de chaque indice est indiquée entre crochets à sa droite :

(2) Dans la matinée, un pont de bateaux est lancé à Drohizyn, 80 km plus au nord. La tête de pont ainsi créée fut appuyée par l'emploi de 80 chars Pz-III submersibles. **La résistance des Soviétiques**$_{\text{[Im:SNdef]}}$ est assez décousue sur la plus grande partie du front. **Elle**$_{\text{[Ia:pro]}}$ est acharnée sur quelques points, comme la citadelle de Brest-Litovsk [...]. Sans appui d'aucune sorte, **les soldats soviétiques**$_{\text{[Im:SNdef]}}$ de la citadelle sont totalement encerclés et sans espoir de secours [...]. **Ils**$_{\text{[Ia:pro]}}$ continuent à se battre en dépit de la disproportion des forces et de l'emploi d'artillerie de siège lourde par les Allemands comme les mortiers de 620 mm. La seule 45e division d'infanterie affectée à la prise de la forteresse déplorera 482 tués (dont 80 officiers) et plus de 1 000 blessés. **Les Russes**$_{\text{[Im:SNdef]}}$ perdront environ 2 000 à 2 500 tués et autant de prisonniers. Mais par son action, **cette résistance**$_{\text{[Ia:SNdem]}}$ ralentit considérablement le mouvement des unités d'infanterie qui doivent empêcher les troupes soviétiques de s'échapper de la poche de Bialystok-Minsk. Pendant ce temps, malgré quelques contre-attaques soviétiques, les unités mécanisées [...]
[wik2_operationBarbarossaCT_coder3_1253783856781]

Dans cet exemple, l'annotateur a reconnu une CT autour du topique de « résistance soviétique ». Cette CT est signalée par six indices. Trois d'entre eux sont issus du prémarquage automatique (indiqués par le label [Ia:xxx] – « a » pour « automatique ») et trois ont été ajoutés manuellement (label [Im:xxx]). Ces indices sont également associés à une catégorie morpho-syntaxique : SN défini (SNdef), pronom (pro) et SN démonstratif (SNdem). Ils ont été définis comme des éléments ayant permis à l'annotateur d'interpréter que l'auteur du texte continue à parler de la même entité de discours, du même thème, signalant ainsi une continuité.

Le fait de ne pas avoir restreint, dans le modèle d'annotation, le type d'élément pouvant jouer le rôle d'indice s'est avéré avantageux. Cela a en effet permis de recueillir des expressions référentielles associées à une large palette de référents (référents humains et non-humains, référents concrets ou abstraits) et rend possible des analyses contrastives du fonctionnement des structures fondées sur la nature ontologique des référents. Cette absence de restriction a également permis la prise en compte de cas particuliers de reprise référentielle généralement exclus des schémas d'annotation portant sur la coréférence, comme par exemple les anaphores résomptives ou les mentions « floues » (Delaborde et Landragin, 2019).

Cependant, même si aucune restriction n'a été explicitement formulée s'agissant de la définition des indices, les règles utilisées pour le prémarquage automatique ont certainement eu une influence dans la conceptualisation de ce que les annotateurs ont considéré comme pouvant jouer le rôle d'indice. Trois facteurs ont certainement eu un impact sur cette conceptualisation : (1) pour répondre à des contraintes de visualisation, un indice pouvait concerner un à trois mots graphiques ; (2) une grande diversité d'expressions ont été prémarquées : les introducteurs de cadre, les connecteurs, les expressions anaphoriques, les amorces d'énumération ; (3) dans la mesure où le prémarquage est automatique, il n'est pas parfait et peut laisser place à des erreurs. Ces trois facteurs ont pu installer un certain flou dans la définition que les annotateurs se sont donnés du concept

d'indice, flou dont l'impact et l'influence est difficile à calculer, mais que cette étude peut aider à apprécier.

2.3 État des lieux des données à disposition

L'annotation a été effectuée par trois étudiants en Licence et Master de Sciences du Langage. 9 textes ont été multi-annotés afin de stabiliser le modèle et calculer l'accord inter-annotateur (cf. Colléter et al., 2012, Ho-Dac & Péry-Woodley, 2014). Suite à cette étape, un ensemble de post-traitements ont été réalisés pour nettoyer, compléter et harmoniser les annotations afin de préparer la mise à disposition de la ressource. Ces post-traitements se sont appuyés en grande partie sur les informations fournies par l'analyseur Talismane (Urieli, 2013) et ont principalement concerné l'annotation des indices et notamment la délimitation (lorsque par exemple un déterminant n'avait pas été inclus dans l'indice), la catégorie grammaticale (lorsqu'elle n'avait pas été indiquée par l'annotateur) et la fonction syntaxique des indices annotés (cf. Ho-Dac & Péry-Woodley, 2014 et Federzoni, 2019). Le tableau 1 donne un aperçu quantitatif des données annotées à disposition.

Tableau 1. Aperçu quantitatif des annotations en CT dans la ressource ANNODIS (I pour indice et I_manuel pour indice ajouté manuellement)**.**

Corpus	# mots	# textes	# CT	# I	% I_manuel
ANNODIS	666 000	87	581	3456	49
GEOP	266 000	32	234	1125	50
LING	169 000	25	87	478	56
WIK2	231 000	30	260	1853	47

Comme l'indique ce tableau, le nombre de CT annoté diffère selon le sous-corpus. La signalisation des CT repose en moyenne sur 6 indices annotés, avec un minimum de 2 et un maximum de 40 indices annotés (dans le corpus WIK2). La part des indices prémarqués dans cette signalisation est importante. Parmi les 3 456 indices annotés, les indices prémarqués représentent à peu près la même proportion que les indices ajoutés manuellement (49% des indices ont été ajoutés manuellement) avec des variations entre sous-corpus – le sous-corpus LING se distingue ainsi des deux autres sous-corpus parce qu'il présente une plus forte part d'indices ajoutés manuellement (56%).

Les post-traitements ont permis d'attribuer une catégorie à la majorité des indices. Le tableau 2 présente la distribution de ces catégories en indiquant pour chacune la proportion que représentent les indices ajoutés manuellement (I_manuel).

Tableau 2. Fréquence et distribution des différentes catégories d'indices annotés dans la ressource ANNODIS (% I_manuel pourcentage par catégorie d'indices ajoutés manuellement).

Catégorie	#	%	% I_manuel
SN définis	1198	35	63
Pronoms	1026	30	29
Noms Propres et toponymes	442	13	45
SN démonstratifs	272	8	27
SN possessifs	182	5	38
SN indéfinis	126	4	98
SN sans déterminant	65	2	85

Autres	145	4	98
Total	3456	100	49

Le tableau 2 montre une répartition des catégories en trois tiers : les pronoms (30%), les SN définis (35%) et les autres catégories réunies. Parmi ces dernières, on notera la présence inhabituelle des SN sans déterminant qui correspondent principalement à des titres de section. La catégorie « Autre » recouvre des éléments pour lesquels le choix de la catégorie n'a pas été stabilisé (SN avec déterminant numéral, verbes, indices aux délimitations floues, etc.). Comme attendu, la proportion d'indices ajoutés manuellement est plus faible pour les catégories d'indices les plus facilement repérables automatiquement que sont les pronoms, les SN possessifs et les SN démonstratifs. Il faut cependant noter qu'environ 30% d'indices appartenant à ces catégories ont été ajoutés manuellement.

3 Une méthode pour confronter les CT et leur signalement aux CR et leurs maillons

L'objectif de notre méthode est de comparer les CT, ainsi que leurs indices tels qu'ils ont été annotés selon le protocole décrit précédemment (annotation naïve selon un modèle qui ne vise pas directement les CR), avec les CR, telles qu'elles sont décrites dans les études disponibles, sachant que nous nous focalisons ici sur les expériences d'annotation portant sur le français. La première étape de la méthode consiste à partir des critères de description des CR afin de les appliquer à la description des données recueillies à propos des CT.

Dans les travaux disponibles, on observe une grande hétérogénéité dans la définition des éléments annotés, que ce soit pour les structures ou leur signalisation. La notion de signalisation n'est généralement pas abordée de front, comme cela a été fait dans le projet ANNODIS, mais s'appuie sur une approche ascendante partant des expressions référentielles qui, une fois détectées, constituent des maillons de CR à partir du moment où une relation de coréférence a été identifiée entre eux. Cette méthode permet ainsi de mettre au jour la structure complète de la CR.

Concernant la définition de la textualité d'une CR, les différences principales concernent la délimitation des CR. Beaucoup de travaux considèrent qu'un nombre minimal de trois maillons est nécessaire pour qu'on puisse parler de CR (e.g. Schnedecker, 2005). En revanche, la taille maximale d'une CR n'est généralement pas précisée. Certains modèles envisagent les chaînes de manière globale et choisissent de regrouper l'ensemble des mentions d'un même référent à l'intérieur d'une même chaîne. Ces modèles se distinguent de ceux qui laissent la possibilité de l'annotation de plusieurs chaînes associées à un même référent, à partir du moment où intervient une redénomination entraînant l'ouverture d'une nouvelle chaîne (Schnedecker, 1997). Concernant les maillons, les plus grandes divergences s'observent concernant le type d'éléments pouvant constituer des maillons et la délimitation de ces derniers.

Bien que la plupart des travaux s'accordent sur une définition minimale des maillons correspondant aux expressions référentielles définies par Charolles (2002) telles que les SN, les noms propres et les pronoms (personnels, démonstratifs, adverbiaux, possessifs, indéfinis, relatifs), une variété d'autres types d'éléments peuvent venir s'ajouter à la liste. Ainsi, Landragin (2011) inclut dans la liste des expressions référentielles les sujets zéro et les pronoms réfléchis et proposent d'intégrer dans cette liste les phénomènes d'accord en genre et nombre des verbes conjugués et des participiaux. Les études diffèrent également concernant la délimitation des maillons : syntagmes nominaux complets, modifieurs/compléments d'un syntagme (cf. plus bas l'exemple (18)).

Malgré ces nombreuses différences, un certain nombre de critères semblent faire consensus pour décrire les CR et les maillons qui les composent. Parmi ces critères, certains

peuvent être retenus pour décrire les CT et les indices de la ressource ANNODIS, comme par exemple :

- la longueur des chaînes en nombre de mots,
- la longueur des chaînes en nombre de maillons,
- la nature des maillons (SN définis, indéfinis, pronoms etc.),
- la fonction syntaxique des maillons (sujet, objet, autre).

On peut ajouter à cette liste une distinction propre à la ressource : la distinction entre indice prémarqué et indice annoté manuellement.

Pour toutes ces caractéristiques, une comparaison a été faite entre chacun des sous-corpus et la position de l'indice dans la CT. Cette comparaison s'appuie sur des tests statistiques permettant d'évaluer si les différences observées d'un corpus à l'autre, d'une position d'indice à une autre, sont significatives.

La fréquence de la catégorie des indices annotés est calculée en suivant l'étiquette attribuée à chaque indice, comme dans l'exemple (3) tirés d'un article de géopolitique intitulé « GÉOÉCONOMIE DU BASSIN CASPIEN » :

(3) Les économies de la région sont ainsi devenues de plus en plus dépendantes de l'aide et de l'investissement étrangers.
L'IDE (investissement direct étranger)$_{[Im:SNdef]}$ est nécessaire pour mettre en valeur les ressources de la région. Pourtant, **il**$_{[Ia:Pro]}$ reste encore très faible. Parmi les pays en transition, l'Asie centrale est le parent pauvre du point de vue de **l'IDE**$_{[Im:SNdef]}$. La BERD a calculé, sur la période 1989-1999, que **l'IDE par habitant**$_{[Im:SNdef]}$ avait été de 668 dollars pour les pays d'Europe centrale et orientale. Pour les pays de la CEI, **ce ratio**$_{[Ia:SNdem]}$ était près de cinq fois inférieur, s'élevant à 140 dollars. Si on excepte le Kazakhstan qui a attiré près de 80 % de **l'IDE**$_{[Im:SNdef]}$ en Asie centrale, **l'IDE**$_{[Ia:SNdef]}$ est inférieur à 50 dollars par habitant. Malgré les hydrocarbures et les métaux, l'Asie centrale n'a reçu que 0,3 % **des IDE investis**$_{[Im:SNindef]}$ dans le monde sur la période 1998-2000 **Ce chiffre**$_{[Ia:SNdem]}$ était nul dix ans plus tôt mais seuls les pays en développement du Pacifique sud ont attiré moins de capitaux que les pays d'Asie centrale sur cette période de trois années.
L'investissement est faible. [...]
[geop_28CT_coder2_1280479728969]

La fonction syntaxique de l'indice est obtenue en se fondant sur l'analyse fournie par Talismane simplifiée ici aux cas suivants : si l'indice est inclus dans un constituant analysé par Talismane comme ayant une fonction de sujet, l'indice est considéré comme ayant la fonction sujet ; si l'indice est inclus dans un constituant analysé par Talismane comme un argument du verbe, l'indice est considéré comme ayant la fonction objet. Dans les autres cas, l'indice est considéré comme ayant la fonction syntaxique « autre ». En appliquant ces règles, le 3^{e} indice de l'exemple (3), à savoir *l'IDE*, lui-même pris dans le SN *le parent pauvre du point de vue de l'IDE* est considéré comme un objet alors que tous les autres indices sont analysés en tant que sujets.

4 Caractéristiques des CT de la ressource ANNODIS

4.1 Signalisation des CT : nombre d'indices par CT selon les sous-corpus

Les CT de la ressource ANNODIS sont majoritairement signalées par au moins trois indices (seules 35 sur les 581 CT annotées ne présentent que deux indices). En cela, les CT ne diffèrent pas des CR (cf. section précédente). Plus précisément, les CT sont en moyenne associées à 6 indices, moyenne qui varie en fonction du genre textuel comme l'illustre la figure (1). Dans cette figure, le nombre moyen d'indices par CT dans les trois sous-corpus est indiqué par des losanges. Les lignes horizontales indiquent le nombre moyen de

maillons par CR selon les genres considérés dans les travaux qui ont étudié la variation du nombre de maillons par CR en fonction du genre textuel et/ou du domaine.

Dans les sous-corpus GEOP et LING, les CT annotées présentent un nombre relativement semblable d'indices (respectivement 4,8 et 5,5). En revanche, dans le sous-corpus WIK2, les CT contiennent beaucoup plus d'indices (7,1). Cette différence entre WIK2 et GEOP se révèle significative, avec une valeur-p de 3.993e-07, alors qu'entre WIK2 et LING, la différence n'est pas significative.

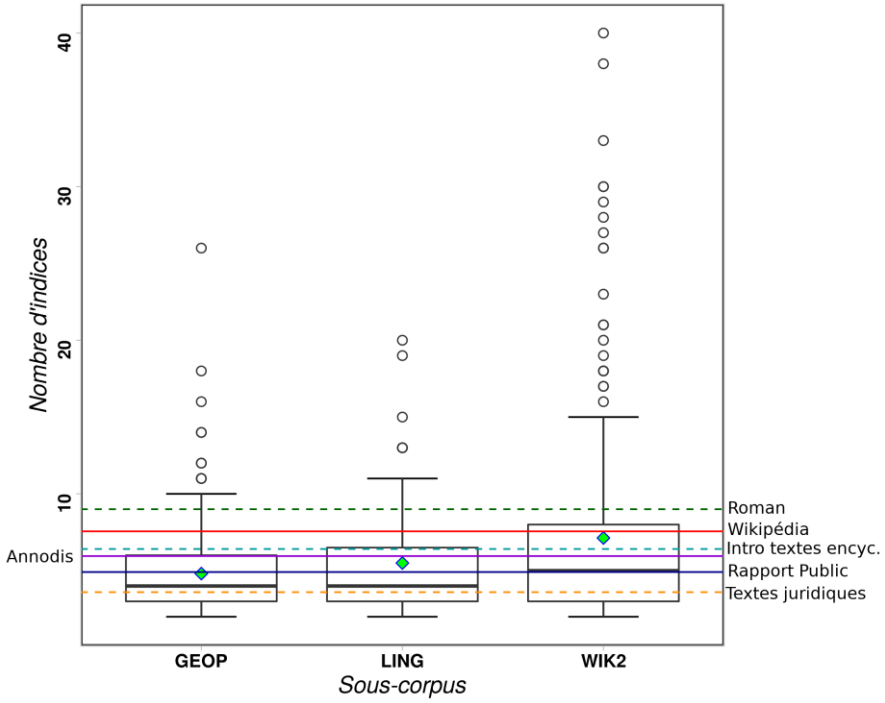


Fig. 1. Nombre moyen d'indices par CT comparé au nombre moyen de maillons par CR dans différents genres textuels et domaines, d'après les travaux de Schnedecker (2014), Longo (2013), Longo & Todirascu (2014) et Todirascu et al., (2017)

Le sous-corpus WIK2 se distingue des deux autres par une plus grande dispersion des valeurs et une plus forte présence de cas « atypiques ». On constate par exemple que parmi les 14 CT contenant plus de 20 indices annotés, 13 sont issues d'articles de Wikipédia et plus précisément d'articles de type biographie. On trouve notamment dans les articles consacrés à Jules César ou à Albert Einstein les 2 CT présentant le plus d'indices annotés, respectivement 40 et 38. Cette différence pourrait signifier que dans les textes encyclopédiques, les CT ont une portée plus longue, comme le montre Schnedecker (2014). Le nombre moyen d'indices signalant une CT (7,1), dans le sous-corpus WIK2, est très proche du nombre moyen de 7,6 maillons par CR observés dans le corpus *Wikipédia* par Todirascu et al. (2017). En revanche, cette moyenne dépasse celle observée par Schnedecker (2014) dans le corpus *Introductions de textes encyclopédiques* (6,4 maillons par CR). Cette différence peut s'expliquer pour deux raisons. Premièrement, Schnedecker (2014) ne traite que les introductions des articles, alors que WIK2 se compose des articles entiers. Deuxièmement, Schnedecker (2005, 2014) considère que toute redénomination entraîne l'ouverture d'une nouvelle chaîne ce qui multiplie le nombre de CR et par conséquent réduit le nombre de maillons par CR.

Les CT du sous-corpus GEOP sont en moyenne associées à 4,8 indices, ce qui correspond au nombre moyen de maillons par CR annotées dans le corpus *Rapport Public* (Longo et Todirascu, 2014), qui est, quant à lui, composé de mesures d'adaptations au

changement climatique adoptées par l'Union Européenne. Ce corpus porte sur un thème global, développé sous différentes facettes (agriculture, gestion des zones maritimes etc.) et il est riche en référents non-humains. Ces caractéristiques le rapprochent du sous-corpus GEOP, dont les textes se situent dans le domaine de la géopolitique étudiant notamment les effets de la géographie (humaine et matérielle) sur la politique internationale et les relations internationales. Il s'agit en particulier d'études régionales, du climat, de la démographie susceptibles d'instancier des référents humains ou non-humains et de porter sur un thème global argumenté sous différents aspects. En revanche, bien qu'il s'agisse des textes non-narratifs, les CR du corpus *Textes juridiques* composés d'arrêtés de la Cour de Justice européenne (Longo et Todirascu, 2014) contiennent moins de maillons (3,5/3,6 maillons) que d'indices pour les CT de GEOP, ce qui semble confirmer qu'en plus du genre textuel, le domaine aussi influence la composition des chaînes.

La comparaison des CT avec les CR annotées dans les romans par Longo (Longo 2013) fait apparaître un nombre moyen de maillons par CR plus élevé que de le nombre d'indices par CT, et ce quel que soit le sous-corpus. Dans l'extrait annoté issu du roman *Les trois mousquetaires*, Longo remarque que ce sont les passages descriptifs qui favorisent la mise en place de CR longues, constat que l'auteur ne fait pas dans des textes non-narratifs.

Au terme de cette comparaison, nous pouvons affirmer que les structures annotées dans ANNODIS correspondent à des CR.

4.2 Catégorisation des indices signalant les CT

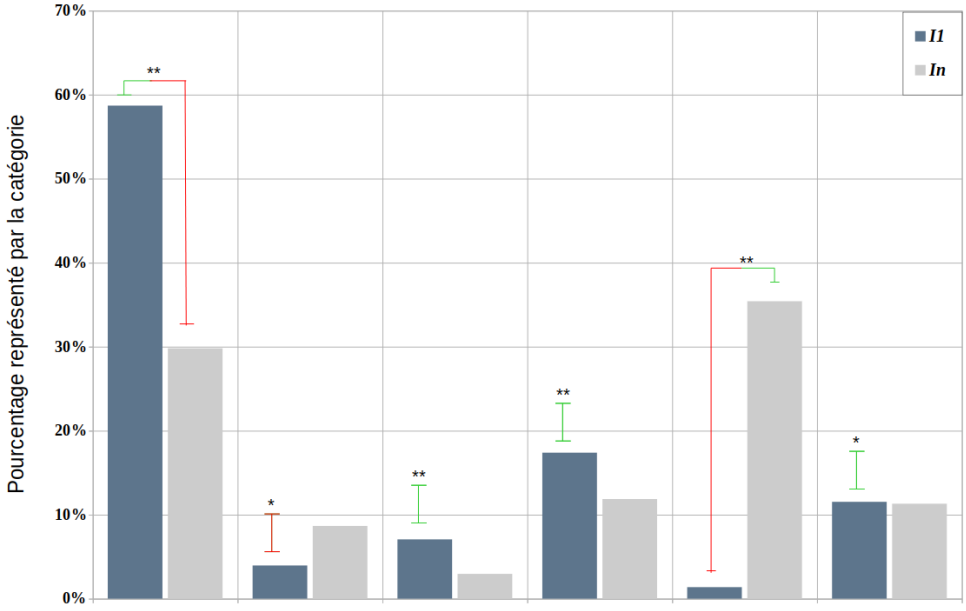
Selon la théorie de l'accessibilité (Ariel, 2001), les expressions référentielles fournissent une information sur le degré d'accessibilité d'un référent à un moment donné du discours. En d'autres termes, elles fournissent les « instructions » dont les interlocuteurs ont besoin afin de récupérer l'information pertinente en mémoire. Suivant l'échelle proposée par Ariel (2001) et adaptée à l'étude du français écrit par Ho-Dac (2007), on peut distinguer 8 degrés d'accessibilité avec les expressions de basse accessibilité, comme les SN définis (degré 0) ou les noms propres nouveaux (degré 1), les expressions de faible accessibilité présentant un degré compris entre 2 et 4 depuis les SN définis jusqu'au noms propres répétés, les expressions d'accessibilité moyenne qui se manifesteraient sous forme de SN démonstratifs (degré 5 - 6) et les expressions de haute accessibilité, tels les pronoms ou les sujets zéros (degré 7). L'application de cette théorie aux structures ayant trait à la continuité référentielle, CT ou CR, devrait prédire un degré d'accessibilité plus faible pour les premiers indices/maillons que pour les indices/maillons suivants, même si certaines limites ont été soulevées dans la littérature, notamment le fait que cette échelle devrait prendre en compte le genre textuel (Schnedecker, 2005).

En appliquant cette échelle aux indices signalant les CT (cf. Tableau 2 *supra*), on distingue d'une part les SN définis, expressions de moyenne accessibilité, représentant 35% des indices, et d'autre part, les pronoms, expressions de la plus haute accessibilité, constituant 30% des indices. Les 35% restants se partagent entre une accessibilité élevée (7,9% de SN démonstratifs et 5% de SN possessifs) et des indices de faible accessibilité (4% de SN indéfinis et 2% de SN sans Déterminant). Pour interpréter les pourcentages obtenus concernant les noms propres (12,8% des indices), une analyse plus fine serait nécessaire pour distinguer les noms propres repris des noms propres nouveaux (première mention). Il en va de même pour les 4% d'indices encore non catégorisés.

Cet état des lieux nous permet de procéder à l'étude de la répartition de ces catégories et degrés selon la position des indices dans les CT et notamment en proposant une première analyse contrastive entre le premier indice (I1), *a priori* moins accessible, et les indices suivants (In).

La figure 2 présente la répartition entre I1 et In des catégories les plus fréquentes : SNdef > Pro > NPP > SNdem > SNind. Les catégories restantes ont été regroupées sous le

label « autre » (SN possessifs, SN sans déterminant et indices pour lesquels le choix de la catégorie n'a pas encore été fait). Pour chacune de ces catégories, le test du Khi2 a été utilisé pour apprécier la force de la relation entre la catégorie et l'indice, en tenant compte de la position de l'indice : I1 ou In. Les relations significatives sont représentées par les barres au-dessus desquelles on a placé les symboles * ou ** pour représenter la force de la relation en fonction du score obtenu au test. Le tableau qui accompagne la figure donne les effectifs pour chacune des catégories dans les deux positions I1 et In.



	SNdef (2-3)	SNdem (5-6)	SNind (0)	NPP (1 ou 4)	Pro (7)	Autre
I1	341	23	41	101	8	67
In	857	249	85	341	1018	325

Fig. 2. Répartition des catégories en fonction de la position des indices : première position (I1) *vs.* autres positions (In). Le degré d'accessibilité selon l'échelle d'Ariel (2001) est indiqué entre parenthèse à côté des catégories.

La figure 2 montre que la catégorie des premiers indices (I1) va dans le sens de la théorie de l'accessibilité. En effet, les I1 sont associés à de faibles degrés d'accessibilité : les SN définis dominent (SNdef : 58,7%), mais on trouve aussi des noms propres (NPP : 17,4%) et des SN indéfinis (SNind : 7,1%). Toutes ces associations sont très significatives, comme l'indiquent les **.

Les indices non initiaux (In) montrent une distribution qui se distingue nettement de la distribution observée pour les I1. La catégorie la plus représentée est celle des pronoms (Pro : 35% des In). La relation significative entre In et la catégorie pronom n'a rien de surprenant dans la mesure où ces expressions de très haute accessibilité présupposent que le référent est le topique.

Concernant les autres catégories associées aux In, seule la relation négative avec les SNdef est significative, même si les SNdef et les NPP restent assez fréquents (32,4% et 30,5% respectivement).

Nos résultats corroborent ceux des travaux concernant les CR (Corblin, 1987, Cornish, 1998, Manuélian, 2003 et Schnedecker et Landragin, 2014). Ils montrent en effet que les SNdef et les NPP, expressions de basse accessibilité ayant la caractéristique d'être complètes sémantiquement et pouvant ainsi être interprétées indépendamment de leur contexte, sont de bons candidats pour constituer le premier indice d'une CT.

La relation significative observée entre SNind et I1 va dans le sens des travaux théoriques de Corblin (1987)[2] et des observations dans le corpus *Textes juridiques* (Longo et Todirascu, 2014). D'autres travaux en corpus ne vont cependant pas dans ce sens, comme ceux de Schnedecker et Landragin (2014) qui affirment que les SNind ne favorisent pas la mise en chaîne car le référent qu'ils introduisent est *a priori* non spécifique, ou spécifique mais indéterminé et ne constituant pas le « focus d'attention ».

Nos résultats montrent enfin que la présence des SNdef et NPP en position In n'est pas à exclure. Ces expressions s'avèrent en effet les plus aptes non seulement à réactiver un référent précédemment mentionné, mais qui ne serait plus activé dans la mémoire du lecteur, mais aussi à éviter une ambiguïté, lorsqu'il y a compétition des référents comme on le voit par exemple en (4) où le SNdef *le président* apporte l'information nécessaire pour que le lecteur interprète l'énoncé :

(4) Au fur et à mesure que l'échéance du 15 janvier 1991 approchait, les positions des uns et des autres se firent de plus en plus nettes, certains n'hésitant pas à évoquer l'impeachment si le président choisissait de ne pas se présenter au Capitole

George H.W.$_{\text{[Ia:NPP]}}$ Bush s'inquiéta de la volonté du régime irakien de se doter d'armes de destruction massive, ce à quoi Al Gore, alors sénateur (démocrate, Tennessee), répliqua qu'il s'agissait d'une manoeuvre maladroite destinée à gagner le soutien du Congrès. Se sentant dans une situation de plus en plus délicate, **le président**$_{\text{[Ia:SNdef]}}$ comprit que la guerre était inévitable, [...]. **Il**$_{\text{[Ia:pro]}}$ estimait en effet que [...]. Ainsi, comme **il**$_{\text{[Ia:pro]}}$ l'écrivait [...] :

« [...].

[geop_19CT_coder1_1253610481687]

D'autres expressions de haute accessibilité peuvent occuper la position In, tels les SNdem, qui, selon Corblin (1987), permettent une reprise immédiate d'un SNind ou une reprise recatégorisante, c'est-à-dire apportant une information nouvelle sur le référent. Les deux cas sont attestés dans nos données. On en a un exemple en (5), où le deuxième indice est un SNdem de reprise immédiate d'un SNind et les autres indices In, également réalisés par le biais de SN démonstratifs – débutant chacun un nouveau paragraphe –, sont des reprises recatégorisantes :

(5) [...] Cette généralisation est déjà entrevue, dès la fin des années 1950, par Benveniste.

Indépendamment des recherches de la philosophie [...], Benveniste développe au cours des année 1950 **une théorie globale sur la langue**$_{\text{[Im:SNind]}}$ et, plus particulièrement sur l'analyse du discours (Dessons, 2006). **Ce nouveau champ**$_{\text{[Ia:SNdem]}}$ d'investigation lui permet d'étudier la façon dont l'homme se projette dans la langue.

Ces analyses$_{\text{[Ia:SNdem]}}$, publiées dans diverses revues, seront reprises dans les Problèmes de linguistique générale [...]

Ce renouveau$_{\text{[Ia:SNdem]}}$ projette la méthode structuraliste en dehors de son cadre traditionnel, [...]

Cet essor$_{\text{[Ia:SNdem]}}$ du structuralisme permet à la linguistique[...]

Ces aspects vont ensuite se développer essentiellement en ...

[ling_poibeauCT_coder2_1322415685992]

Certains de nos résultats sont plus surprenants. On observe notamment la présence en I1 de SNdem et de pronoms, position dont ils sont théoriquement exclus : les SNdem, dans la mesure où ils coréfèrent à une entité déjà mentionnée, et, les pronoms, en raison de leur « incomplétude sémantico-pragmatique » (cf. Corblin 1987 ; Cornish 1998 ; Manuélian 2003).

23 SNdem se trouvent en position de premier indice. Deux raisons peuvent être invoquées pour expliquer ce constat. D'abord, les SN démonstratifs permettent des emplois déictiques référant à l'article lui-même ou à l'étude qui y est présentée, comme on le voit en (6) :

(6)Introduction

L'objectif de **cette étude**$_{[Im:SNdem]}$ est double. Il s'agit de […] **Un tel travail**$_{[Im:SNind]}$ devrait [...] **Cette étude**$_{[Ia:SNdem]}$ permettra [...]
[geop_19CT_coder1_1253609208609]

Ensuite, nous observons des cas d'encapsulation où le SNdem résume un segment de texte comme en (7) :

(7) Les deux principaux groupements d'intérêt concernés et menacés [...] ne se sont pas laissé imposer séparément un agenda […]. Ils ont au contraire annoncé, en 1996, qu'ils feraient des propositions communes pour [...]. D'autres accords ont aussi été conclus avec les constructeurs automobiles japonais et coréens.
Malgré les appréhensions de beaucoup de députés européens, **cette approche**$_{[Ia:SNdem]}$ a été acceptée et le premier programme Auto-Oil voté [...]
Ce processus$_{[Ia:SNdem]}$ [...] **Il**$_{[Ia:pro]}$ [...]
[geop_15CT_coder1_1255361409500]

Pour ce qui concerne la présence des pronoms en I1 de CT (8 cas), nous avons remarqué qu'il s'agit le plus souvent d'une « erreur » dans le sens où le segment CT annoté ne contient pas la première mention du référent, comme c'est le cas dans l'exemple (8), où la phrase contenant l'antécédent du pronom, *Léonard*, ne fait pas partie de la CT :

(8) De plus, dans ces épures, Léonard ne pose jamais le problème de la force motrice.
Dans une lettre adressée à Ludovic Sforza, **il**$_{[Ia:Pro]}$ prétend être capable de construire toutes sortes de machines à la fois pour la protection de la ville et pour le siège. Quand il a fui à Venise en 1499, **il**$_{[Ia:Pro]}$ a trouvé un emploi d'ingénieur et a développé un système de barrières mobiles [...]. **Il**$_{[Ia:Pro]}$ a également eu pour projet de détourner la circulation de l'Arno [...].
Ses carnets [...]
[wik2_leonardDeVinciCT_coder2_1257956898336]

On peut interpréter cette erreur d'une part comme une conséquence de l'approche descendante et d'autre part comme une conséquence de la consigne d'annotation. Nous avons en effet indiqué que les annotateurs des CT, qui ne sont pas des experts de la coréférence, n'ont pas procédé à une lecture linéaire et exhaustive du texte. Au contraire, ils ont parcouru le texte « de haut », en réalisant des zooms sur les zones présentant plusieurs indices prémarqués. De plus, comme nous l'avons dit plus haut, aucune consigne relative à l'ouverture d'une CT n'a été fournie. Ainsi, dans l'exemple (8), la série de pronoms prémarqués [Ia:Pro] se trouve concentrée dans un paragraphe dont la phrase incluant l'antécédent est exclue[3], l'annotateur n'ayant pas cherché à d'identifier cette première mention susceptible d'ouvrir la chaîne. Dans cet exemple, on peut penser que le saut de paragraphe a joué un rôle décisif dans la délimitation du segment, en ce sens qu'il signale une discontinuité dans le discours.

Ces exemples nous permettent d'attirer l'attention sur un facteur peu pris en compte dans les travaux portant sur les CR et que l'annotation descendante semble favoriser. Il s'agit du fait que, pour annoter, les annotateurs s'appuient sur la structure du texte et notamment sur le rôle de la segmentation en paragraphes et des titres dans le processus d'interprétation du texte.

4.3 Fonction syntaxique des indices

Comme indiqué dans la partie 3, chaque indice est accompagné de sa fonction syntaxique. Dans cette étude, seules les fonctions de sujet et d'objet ont été considérées, les autres fonctions étant indifféremment regroupées dans une même catégorie appelée « autre ».

La fonction la plus fréquente des indices annotés est la fonction de sujet. Elle concerne près de 3/4 des indices, soit 71,4% (avec un minimum de 67,4%, dans LING, et un

maximum de 72,2%, dans GEOP). Ce résultat va dans le sens à la fois des travaux concernant la structure informationnelle qui considèrent que la fonction sujet serait la position privilégiée de l'expression du topique (on parle de « centre préféré », cf. Cornish, 1998, 2000) et des résultats de l'annotation des CR dans les introductions des textes encyclopédiques décrite dans Schnedecker (2005).

Ce résultat doit être mis en relation avec deux spécificités de la ressource ANNODIS. Premièrement, le prémarquage a concerné exclusivement les indices en position de sujet. Deuxièmement, la consigne donnée aux annotateurs indiquait explicitement que « l'expression [du] référent commun doit passer nécessairement par le sujet grammatical » (Colléter et al., 2012). Malgré ces spécificités qui pouvaient amener l'annotateur à ne considérer que les indices sujets, on observe que près de la moitié des indices ajoutés manuellement (48,5%) ne sont pas en position sujet.

Si l'on observe maintenant les indices en position d'objet, on note des cas de transitions de fonction syntaxique comme en (9) où I1 en position objet est repris dans l'énoncé suivant avec un I2 en position de sujet.

(9) Le parti minoritaire dispose d'**un représentant**$_{[Im:SNind]}$ qui se fait l'écho des membres de son parti. Dans un système partisan, **ce représentant**$_{[Ia:SNdem]}$ ne peut qu'exprimer [...]. Cependant [...] **il**$_{[Ia:Pro]}$ [...]. **Ce personnage**$_{[Ia:SNdem]}$ [...]
[geop_19CT_coder1_1253613767234]

Cet exemple de transition de fonction syntaxique a également été observé à propos des CR par Todirascu et al. (2017). Une modélisation de ces transitions est proposée dans la théorie du centrage (Walker et al., 1998) qui vise à prédire le choix d'une expression référentielle en fonction de la présence, la position et la fonction syntaxique de son référent dans l'énoncé précédent. Dans cette théorie, les auteurs distinguent deux types de transition entre expressions coréférentielles[4] : la transition par continuation qui concerne les situations où les deux expressions coréférentielles, In et In+1, sont les centres préférés (i.e. sujets) et la transition par rétention dans laquelle seul In+1 est en position sujet. La transition par continuation est la transition à préférer pour maintenir la continuité référentielle.

En nous fondant sur ces propositions, nous avons mené une étude exploratoire des transitions de fonction syntaxique entre I1 et le deuxième indice I2 des CT en ne considérant uniquement les indices en position sujet ou objet (cf. tableau 3).

Tableau 3. Fréquence et distribution des transitions entre I1 et I2 selon leur fonction syntaxique

{I1 sujet – I2 sujet}	{I1 objet – I2 sujet}	{I1 objet – I2 objet}	{I1 sujet – I2 objet}
285 (49 %)	99 (17%)	26 (4 %)	31 (5%)

Les paires {I1 sujet – I2 sujet} relevant de la transition par continuation sont les plus fréquentes (49%). Ces paires se distinguent nettement des paires {I1 objet – I2 sujet} relevant quant à elles de la transition par rétention. Dans la mesure où les relations entre les indices qui la signalent sont davantage des transitions par continuation, les CT annotées correspondent principalement à des continuités.

Conclusion et perspectives

Au terme des analyses présentées ici nous nous proposons de dresser un bilan en deux temps : le premier temps est consacré aux similarités entre les chaînes de référence et les chaînes topicales annotées dans la ressource ANNODIS, le second temps concerne les spécificités de ces structures et leur intérêt pour l'étude de l'organisation du discours.

Que ce soit en terme de signalisation (nombre d'indices), de variation en fonction des genres, de nature morpho-syntaxique et de fonction syntaxique des expressions jouant le rôle d'indice, les chaînes topicales apparaissent très comparables aux chaînes de référence

annotées dans les autres ressources disponibles pour le français. Les différents points de sondage décrits dans la présente étude soulignent d'autres similarités comme par exemple le fait que la redénomination d'un nom propre peut instruire l'ouverture d'une nouvelle chaîne (Schnedecker, 1997) ou le fait que les SN démonstratifs ne peuvent pas se trouver en première position dans une chaîne (Corblin, 1987 ; Manuélian, 2003). Ces résultats nous confortent dans l'idée qu'une annotation de type descendante et non experte fournit des données exploitables pour l'étude de la continuité référentielle. Au vu de ces résultats, il semble permis d'envisager des campagnes d'annotation à grande échelle portant sur des textes longs et relevant de genres textuels variés.

L'étude a également mis en évidence un certain nombre de spécificités des chaînes topicales de la ressource ANNODIS. Du fait de la méthode d'annotation appliquée, la ressource offre un regard inédit sur l'organisation topicale des textes et sur les processus mis en place pour détecter la continuité référentielle. Les exemples discutés ici donnent une idée de la « liberté » qu'ont pu prendre les annotateurs pour définir le début d'une chaîne topicale et pour sélectionner les indices susceptibles de la signaler. Les chaînes annotées donnent également accès aux processus d'interprétation mis en œuvre par les annotateurs non experts pour identifier les (dis)continuités et devraient permettre d'apporter des réponses à certaines questions, comme par exemple : quels indices signalent le début d'une chaîne ? Quelle est la particularité des SN démonstratifs dans l'ouverture et le maintien des continuités référentielles ? Quel est le rôle des titres et de la segmentation en paragraphes dans la délimitation des chaînes ? Répondre à ces questions permettrait de mettre au jour les configurations d'indices qui contribuent à la signalisation et à l'interprétation des continuités référentielles, à la manière dont cela a été fait notamment par Péry-Woodley et al. (2017) et Rebeyrolle et Péry-Woodley (2014) pour les structures énumératives.

---

[1] Dans tous les exemples, les segments de texte correspondant à une CT annotée sont surlignées en gris. Pour des raisons de place, le texte des exemples est tronqué par endroits ([...]). L'exemple complet est accessible via le site de la ressource en utilisant l'identifiant qui suit le texte de l'exemple. Les sauts de paragraphe sont indiqués par un retour à la ligne.

[2] Corbin (1987) considère que les syntagmes nominaux indéfinis peuvent occuper uniquement la position de premier maillon parce qu'ils jouent le rôle d'introducteur des nouveaux référents.

[3] Cette série de pronoms sujets prémarqués peut également expliquer la non annotation du pronom *il* dans la subordonnée temporelle *Quand il a fui à Venise*.

[4] La théorie du centrage considère quatre types de transition possibles entre deux énoncés dont seuls deux concernent des transitions entre expressions coréférentielles. Les deux autres transitions (déplacement en douceur et déplacement brutal) relèvent de cas où un nouveau référent est introduit.

Références bibliographiques

Ariel, M. (2001). Accessibility theory: An overview. Text representation: Linguistic and psycholinguistic aspects, 8, 29-87.

Asher, N. & Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Asher, N., Muller, P., Bras, M., Ho-Dac, L.-M., Benamara, F., Afantenos, S. & Vieu, L. (2017). ANNODIS and related projects: case studies on the annotation of discourse structure. In Nancy Ide (eds.), *Handbook of Linguistic Annotation* (pp.1241-1264). Springer Netherlands.

Carlson, L., Marcu, D. & Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. *Discourse and Dialogue*, 85-112.

Charolles, M. (2002). *La référence et les expressions référentielles en français*. Editions Ophrys.

Colléter, M., Fabre, C., Ho-Dac, L.-M., Péry-Woodley, M.-P., Rebeyrolle, J. & Tanguy, L. (2012). La ressource ANNODIS multi-échelle : guide d'annotation et bonus (Rapport technique).

Corblin, F. (1987). *Indéfini, défini et démonstratif*. Droz. Genève.

Cornish, F. (1998). Les chaînes topicales : leur rôle dans la gestion et la structuration du discours. *Cahiers de grammaire*, 23, 1, 9-40.

Cornish, F. (2000). L'accessibilité cognitive des référents, le centrage d'attention, et la structuration du discours : une vue d'ensemble. *Verbum*, 22, 1, 7-30.

Delaborde, M. & Landragin, F. (2019). En quoi le pronom *on* a-t-il une valeur anaphorique ?, *Cahiers de praxématique*, 72.

Federzoni, S. (2019). *Évaluation et exploitation de la Ressource ANNODIS pour la détection des chaînes de référence*. Mémoire de Master 2.

Ho-Dac, L.-M. (2007). *La position initiale dans l'organisation du discours : une exploration en corpus*. Thèse de doctorat non publiée. Université Toulouse le Mirail-Toulouse II.

Ho-Dac, L.-M. & Péry-Woodley, M.-P. (2014). Annotation des structures discursives : l'expérience ANNODIS. In 4e *Congrès Mondial de Linguistique Française* (*CMLF 2014*), pp.2647-2661.

Lambrecht K. (1994) *Information structure and sentence form. Topic focus and the mental representation of discourse referents*. Cambridge University Press: Cambridge, Massachusset.

Landragin, F. (2011). Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. *Corpus*, 10, 61-80.

Landragin, F. (2015). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'AFIA*, 92, 11-15.

Longo, L. (2013). *Vers des moteurs de recherche "intelligents" : un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence*. Thèse de doctorat réalisée dans le cadre d'une CIFRE (convention industrielle de formation par la recherche) avec la société RBS (Ready Business System). Université de Strasbourg.

Longo, L., & Todirascu, A. (2014). Vers une typologie des chaînes de référence dans des textes administratifs et juridiques. *Langages*, 195 (3), 79-98.

Mann, W. C. & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8 (3), 243-281.

Manuélian, H. (2003). *Descriptions définies et démonstratives : analyses de corpus pour la génération de textes*. Thèse de doctorat non publiée. Université de Nancy 2.

Péry-Woodley, M.-P., Afantenos, S., Ho-Dac, L.-M. & Asher, N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Traitement Automatique des Langues*, 52 (3), 71-101.

Péry-Woodley, M.-P., Ho-Dac, L.-M., Rebeyrolle, J., Tanguy, L. & Fabre, C. (2017). A corpus-driven approach to discourse organisation: from cues to complex markers. *Dialogue & Discourse*, 8, 66 - 105.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K. & Webber, B. L. (2008). The penn discourse treebank 2.0. In *Proceedings of the Sixth International Language Resources and Evaluation* (*LREC'08*), Marrakech, Morocco.

Rebeyrolle, J. & Péry-Woodley, M.-P. (2014). Énumération et structuration discursive. In *Actes du 4e Congrès Mondial de Linguistique Française*, Berlin, Allemagne, pp.3183-3196.

Schnedecker, C. (1997). *Nom propre et chaînes de référence*. Librairie Klincksieck.

Schnedecker, C. (2005). Les chaînes de référence dans les portraits journalistiques : éléments de description. *Travaux de linguistique*, 2, 85-133.

Schnedecker, C. (2014). Chaînes de référence et variations selon le genre. *Langages*, 3, 23-42.

Schnedecker, C. & Landragin, F. (2014). Les chaînes de référence : présentation. *Langages*, 195, 3-22.

Todirascu, A., François, T., Bernhard, D., Gala, N., Ligozat, A.-L. & Khobzi, R. (2017). Chaînes de référence et lisibilité des textes : Le projet ALLuSIF. *Langue française*, 195 (3), 35-52.

Urieli, A. (2013). Robust french syntax analysis: reconciling statistical methods and linguistic knowledge in the talismane toolkit. Thèse de doctorat non publiée. Université de Toulouse II le Mirail.

Walker, M. A., Joshi, A. K. & Prince, E. F. (1998). *Centering theory in discourse*. Oxford University Press.

Webber, B., Egg, M. & Kordoni, V. (2012). Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

Widlöcher, A., & Mathet, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles* (*TALN'09*), session posters, Juin 2009, Senlis, France.