Université de Strasbourg

UFR des Lettres

Travail d'Étude et de Recherche

Master en Sciences du Langage
(2015–2017)

Étude des chaînes de référence dans les articles de recherche de format IMRaD :

problèmes d'annotation, analyse quantitative et qualitative

rédigé par Bruno Oberlé

sous la direction de Mme Catherine Schnedecker

soutenu devant un jury composé de :

Mme Amalia Todirascu	LiLPa, Université de Strasbourg
Mme Catherine Schnedecker	LiLPa, Université de Strasbourg

Sommaire

Abréviations et symboles

1 Indications sur les maillons

Nous délimitons les maillons des chaînes par des crochets. L'index associé indique la chaîne à laquelle le maillon appartient :

[Le petit chat de [la voisine]$_j$]$_i$ joue.

Les exemples authentiques, tirés des textes du corpus, sont suivis du numéro du texte (T0 à T4) :

Le Groupe des Treilles est un complexe archéologique... (T1)

Les exemples qui ne sont pas suivis d'un numéro de texte sont fabriqués :

Le petit chat de la voisine est tout petit.

Le symbole « ø » (ensemble vide) représente un pronom zéro.

2 Colonnes des tableaux de données statistiques

- L CR : longueur moyenne des chaînes (en maillons),
- DI : distance intermaillonnaire (en tokens),
- L MA : longueur moyenne des maillons (en tokens),
- DENS : densité globale (en maillons par paragraphe),
- CNSF : coefficient normalisé de stabilité formelle,
- CNSL : coefficient normalisé de stabilité lexicale,
- PAR : nombre de paragraphes.

Le mode de calcul de tous ces indicateurs sont expliqués dans le corps du texte.

Introduction

Tout texte parle de *quelque chose*. Il introduit *cette chose*, *la* décrit, puis passe à autre chose, avant, peut-être, d'*y* revenir, et d'*en* parler plus avant. Pour ce faire, un texte utilise différentes expressions linguistiques (comme « quelque chose », « cette chose », « la », « y » ou encore « en »), qui toutes renvoient à la *même* chose. C'est la séquence de ces termes, et la relation qui les unit, qui feront l'objet de notre étude.

1 Les chaînes de référence

« La suite des expressions d'un texte entre lesquelles l'interprétation construit une relation d'identité référentielle » est ce que l'on appelle une *chaîne de référence* (Corblin, 1985a, p. 123). Les expressions linguistiques qui la composent sont des *maillons*. Une chaîne est constituée d'au moins *trois maillons* (Schnedecker, 1997, pp. 8–10), sinon la notion n'est guère pertinente, et les concepts de « coréférence » et d'« anaphore » suffisent à caractériser le phénomène.

Illustrons cette définition par le texte suivant (extrait des fables d'Ésope), où l'on voit deux chaînes : celle, indexée i, dont le référent est « le laboureur », et celle, indexée j, dont le référent est « l'aigle » :

(1) [Un laboureur]$_i$, ayant trouvé [un aigle pris au filet]$_j$, fut si frappé de [sa]$_j$ beauté qu'[il]$_i$ [le]$_j$ délivra et [lui]$_j$ donna la liberté. [L'aigle]$_j$ ne [se]$_j$ montra pas ingrat envers [[son]$_j$ bienfaiteur]$_i$...

Outre la perspective linguistique, qui nous intéresse ici et que nous développerons ci-dessous, les chaînes de référence ont surtout été étudiées dans une perspective référentielle, tant du point de vue philosophique (Chastain, 1975) que du point de vue linguistique (Corblin, 1985a, 1987, 1995). Elles ont aussi été étudiées du côté du traitement automatique des langues (TAL) ; Schnedecker et Landragin (2014) expliquent que les recherches portent surtout sur la détection automatique des relations anaphoriques et coréférentielles, notamment en ce qui concerne les redénominations par des entités nommées. Seule Longo (2013) (voir aussi Longo et Todirascu (2015)) cherche à identifier automatiquement les chaînes en déterminant les référents saillants dans le cadre de l'échelle d'accessibilité d'Ariel (1990).

Par ailleurs, deux projets se sont attachés à la description des chaînes de référence : MC4 (« Modélisation Contrastive et Computationnelle des Chaînes de Coréférence »), synthétisé par Schnedecker et Landragin (2014) et Mélanie-Becquet et Landragin (2014) ; et Democrat (« DEscription et MOdélisation des Chaînes de Référence : outils pour l'Annotation de corpus et le Traitement automatique »), que nous décrirons plus avant ci-dessous.

Dans ce travail, nous serons surtout concerné par l'aspect linguistique, dont l'étude a été initiée par Schnedecker (1997). Nous en présenterons les principales questions, synthétisées par Schnedecker et Landragin (2014).

Qu'est-ce qu'un maillon ? Pour Charolles (1988, p. 8), « seules peuvent appartenir (donner lieu à) une chaîne des expressions employées référentiellement, c'est-à-dire toutes et rien que les expressions nominales (ou pronominales) permettant d'identifier un individu (un objet de discours) quelle que soit sa forme d'existence (personne humaine, événement, entité abstraite) ». On peut pourtant se demander s'il ne faudrait pas aussi annoter des prédicats, qu'ils soient des noms ou des verbes, comme le proposent Longo et Todirascu (2014).

Il faut aussi examiner le cas des « éléments non-référentiels qui participent… à la coréférence » (Landragin, 2011, p. 63), comme les sujets zéro ou les marques d'accord. Landragin, par exemple, propose d'annoter certains de ces éléments (qu'il nomme « maillons faibles »), mais pas tous.

Quelles sont les limites des chaînes ? On peut s'interroger sur la persistance d'une chaîne sur tout un texte : un personnage d'un roman initie-t-il une chaîne qui s'étend sur *tout* le roman ? On peut trouver des critères textuels pour couper les chaînes, par exemple lorsqu'on trouve un inter-titre, ou bien lorsqu'on commence un nouveau chapitre, voire un nouveau paragraphe (Schnedecker et Landragin, 2014, pp. 5–8). On peut également envisager des critères linguistiques, propres aux chaînes, pour effectuer le découpage, ce que Schnedecker (1997, p. 24) appelle des « bornes internes aux chaînes ». Par exemple, quand une chaîne se compose essentiellement de pronoms personnels, et de quelques expressions nominales, comme dans :

(2) SN1… Pro… Pro… Pro… (SN2)… Pro… Pro… Pro… SN1'… Pro… Pro…

on peut dire que cette

> redénomination est décisive car, en présentant le référent comme il l'était en première mention, elle indique qu'il y a *re-saisie*, *re-démarrage référentiels*. Partant, on dispose de moyen de fractionner de manière très cohérente, les chaînages jusque-là considérés comme des suites non délimitées de maillons (Schnedecker, 1997, p. 32).

Quelle relation établir entre les maillons ? Toutes les relations entre les maillons d'une même chaîne ne peuvent pas être mises sur le même plan. La différence la plus évidente est celle entre anaphore et co-référence ; Corblin (1995, pp. 167–169) en fait deux types de chaînes : les chaînes-A (pour anaphoriques) et les chaînes-R (pour référentielles). L'anaphore concerne « des termes dont l'interprétation n'est pas fixée pour tout emploi » (p. 32), comme le pronom de troisième personne *il* dans :

(3) a. Pierre est venu. *Il* repartira demain. (nous soulignons)

L'anaphore entraîne des « liens anaphoriques », pour lesquels « la connexion à une expression antérieure est déclenchée et régie par le contenu linguistique de la forme » (p. 168). À

l'inverse, dans la coréférence, l'interprétation « est indépendante de son contexte immédiat d'usage » (p. 32), comme le nom propre dans :

(3) b. Pierre est venu. *Pierre* repartira demain. (nous soulignons)

La coréférence entraîne des « liens référentiels », « obtenus sur la base d'un savoir concernant la communication : connaissance directe de l'univers de référence, mémoire d'usage antérieur, hypothèses concernant les intentions du locuteur, etc. » (p. 168).

On peut aussi opposer les liens entre les maillons en regardant le type d'anaphore (Schnedecker et Landragin, 2014) : anaphores fidèles (*un homme... cet homme*), anaphores infidèles hyperonymiques (*un boeuf... l'animal*), recatégorisantes (*Paul... cette andouille*), pronominales (*Paul... il*).

Comment caractériser les chaînes ? Schnedecker (1997, 2005) propose un certain nombre de paramètres des chaînes, résumés par Schnedecker et Landragin (2014), par exemple :

- leur longueur (*i.e.* le nombre de leurs maillons),
- leur portée (sont-elles locales, *i.e.* dépassent-elles, par exemple, le pagaraphe, ou bien globales, *i.e.* couvrent-elles tout le texte, ou du moins une bonne partie ?) (voir Schnedecker, 2014),
- la distance intermaillonnaire moyenne (il y a différentes façons de calculer cette distance : caractères/mots entre les maillons/têtes de maillon),
- la catégorie et la fonction des maillons, notamment du premier maillon de la chaîne,
- le patron, c'est-à-dire la ou les séquences les plus courantes des catégories grammaticales des maillons (voir notamment Schnedecker et Longo, 2012),
- le mode de cohabitation : succession, entrecroisement, dérivation, partition, fusion, déroulement parallèle.

Qu'est-ce que la « référence » ? Les chaînes de référence ont surtout été étudiées lorsque leur référent est humain. Certains auteurs décident ainsi de n'annoter que les référents humains (Landragin, 2011 ; Glikman, Guillot-Barbance et Obry, 2014), ou, lorsqu'ils incluent les chaînes non-humaines, il s'agit généralement de référents concrets ou bien définis (par exemple, l'objet du litige opposant deux parties dans les *year books* étudiés par Capin (2014)). Seules Longo et Todirascu (2014) s'aventurent à annoter des référents abstraits ou prédicatifs, comme la réduction des gaz à effet de serre.

Le genre a-t-il une influence sur le comportement des chaînes ? De nombreuses études montrent que le genre influe grandement sur plusieurs paramètres des chaînes ; voir par exemple Tutin (2002) pour l'influence du genre sur la résolution des anaphores, Longo et Todirascu (2013) pour un aperçu général sur le genre et les chaînes de référence, Schnedecker (2005) pour le portrait journalistique, Schnedecker et Longo (2012) pour les faits divers, Schnedecker (2014) pour les recettes de cuisine et les introductions d'articles d'encyclopédie, Longo et Todirascu (2014) pour les textes administratifs.

Tutin (2002) s'intéresse à la résolution d'anaphores dans un corpus contenant des articles scientifiques, mais sans aborder la notion de *chaîne de référence*. À notre connaissance, donc, aucune étude n'a étudié les chaînes de référence dans les articles de recherche scientifique. C'est ce que nous proposons de faire.

2 Textes scientifiques et intérêt du format IMRaD

Le discours scientifique a fait l'objet de nombreuses études (voir l'état des lieux de Rinck (2010)). Défini comme « un discours produit dans le cadre de l'activité de recherche à des fins de construction et de diffusion du savoir » (Rinck, 2010), il a notamment été étudié dans une perspective linguistique par Poudat (2006), Rinck (2006) et Tutin et Grossmann (2013). Dans le monde anglophone, c'est l'aspect didactique qui est privilégié (Swales, 1990, 2004), par exemple dans le courant *English for Specific Purpose* , un « courant rhétorique qui s'intéresse essentiellement aux descriptions macro-textuelles et à la description des genres d'un point de vue phrastique ou propositionnel » (Poudat, 2006).

Un discours, c'est un ensemble « de pratiques langagières verbales, conçues comme des pratiques sociales,... [qui] recouvre un ensemble de genres » (Rinck, 2006). Les genres du discours scientifiques sont nombreux : Swales (2004) en définit « les territoires » pour en dresser « la carte », il établit des « hiérarchies » de genres, des « groupes », des « réseaux » ou même des « chaînes », par exemple celle qui va de l'appel à communication jusqu'à la présentation orale dans un colloque. Nous nous intéresserons ici à l'article de recherche, compris comme l'exposé d'une étude réalisée par un ou plusieurs chercheurs à destination d'autres chercheurs. Il s'agit en effet du genre le plus saillant et le plus fréquemment étudié (Swales, 1990, 2004, Poudat, 2006, Rinck, 2006), mais jamais du point du vue des chaînes de référence. De plus, certains articles présentent un format très codifié, le format IMRaD.

« IMRaD » est un acronyme anglais pour « *Introduction, Methods, Results and Discussion* » [1]. Ces termes reprennent les inter-titres de certains articles de recherche scientifique qui relatent une expérimentation, d'abord en l'inscrivant dans un cadre de recherche (introduction), puis en décrivant le déroulement de l'expérimentation (méthodologie) et les observations qui en ont été faites (résultats), avant d'en tirer les conclusions (discussion).

Ce format est à lui-seul un objet de recherche linguistique, notamment après le travail séminal de Swales (1990). Pour l'étude des chaînes de référence, il a cependant un autre avantage : une très forte codification.

Tous les auteurs semblent en effet unanimes : « Le format intitulé IMRD ou IMRAD... fournit un cadre routinisé pour l'écriture des articles de recherche » (Rinck, 2010). Milard (2007), qui a interrogé des chimistes (la chimie est une discipline où le format IMRAD est imposé) sur leurs pratiques d'écriture, confirme que « telle qu'elle m'est racontée par les chercheurs, la mise en forme des articles laisse entendre une grande part de routines... Certaines se manifestent par de l'implicite ou des habitudes, des automatismes. »

C'est que les consignes aux auteurs, fournies soit directement par les éditeurs des revues, soit par l'intermédiaire de manuels de style publiés par des associations disciplinaires, sont claires et contraignantes : « *Recent versions of the Publication Manual [of American Psychological Association], filled with detailed prescriptions, convey the impression that writing is primarily a matter of applying established rules* » (Bazerman, 1988, p. 259), si bien que « *the experimental report is to*

[1] On trouve dans la littérature plusieurs abréviations. En plus de « IMRaD », il existe « IMReD » pour la traduction française (le « e » vient de *et*), ou plus simplement « IMRD ». Nous gardons la première forme, qui nous paraît être la plus fréquente, même dans la littéture francophone (par exemple Poudat, 2006 ou Pontille, 2007), et même si Swales préfère, lui, la forme « IMRD ».

have the specified sections: title, abstract, introduction, method, results, and discussion. Each of the last three sections is to be so titled. Each section must conform to detailed instructions, at times resembling a questionnaire in specificity » (p. 260).

Ce que Bazerman semble regretter, c'est l'effacement de l'auteur (Pontille, 2007) ou de l'énonciation (Grossmann, 2012). Mais nous pouvons, pour notre étude des chaînes de référence, tirer parti de cette très forte codification, puisque cela laisse supposer qu'il y aura peu de différence entre les auteurs, et partant entre les textes qui respectent le format IMRaD. De ce fait, nous contrôlons la variable « style de l'auteur » en la réduisant au maximum.

Pourtant, Fløttum (2006a, p. 37), trouve un nombre relativement élevé de différences linguistiques entre les articles de son corpus (qui ne sont pas tous, cependant, au format IMRaD). Elle insiste en conséquence sur le fait qu'il n'y a pas d'article scientifique prototypique, mais qu'il convient plutôt de raisonner en termes de traits linguistiques (par exemple la présence de la première personne) qui sont prototypiques des articles de recherche.

Par ailleurs, si cette routinisation renforce l'homogénéité des textes du corpus, ou du moins de certains de leurs traits linguistiques, elle permet, de plus, de faire correspondre chacune des parties de l'ensemble du corpus, par exemple toutes les introductions, toutes les parties « méthodologie », etc., qui deviennent alors *comparables*, comme le dit Régent (1980) : « Ce cadrage... délimite de vastes unités à l'intérieur desquelles il est possible de dégager des *séquences discursives* comparables, même si le contenu de chaque section n'est pas strictement défini. » Des articles qui ne respectent pas ce format seraient ainsi plus difficiles à analyser en termes de « parties ». C'est le problème qu'a rencontré Poudat (2006, p. 183) avec les articles de linguistique :

> L'article scientifique de linguistique n'étant pas soumis à la structure IMRAD, l'analyse de ses sections est délicate : il serait ainsi peu envisageable d'analyser les « troisièmes sections d'articles » d'une part parce que tous les textes n'en sont pas pourvus, et d'autre part parce qu'elles ne remplissent *a priori* aucune fonction particulière.

En effet, chacune des parties du format IMRaD remplit une fonction particulière : « *the linear structuring of the text in sections imposed by the IMRAD format corresponds to a linear distribution of rhetorical functions across the research article* » (Müller-Gjesdal, 2013). Ces fonctions sont tellement spécifiques, que « *although problem, method, and results must correlate at some level, the author escapes the need for transitions to demonstrate the coherence of the enterprise* » (Bazerman, 1988, p. 260). C'est cela qui permet de les opposer, comme le suppose Swales (1990, p. 177) : « *the more likely we will find that different sections will have different rhetorical features* (e.g. *Introductions in contrast to Methods)* ».

3 Problématique

Nous venons de voir que l'étude des chaînes de référence représentait un intérêt linguistique majeur, mais que leur analyse posait nombre de questions qui n'ont pas eu, jusque-là, de réponses définitives. Par ailleurs, le format IMRaD est un format spécifique de certains articles de recherche, qui laisse supposer une grande homogénéité, au moins pour certains traits lin-

guistiques, entre les textes. Il segmente le texte en plusieurs sections qui remplissent chacune un rôle particulier, et qu'on pourrait opposer à l'aide de certains traits linguistiques.

Nous nous proposons donc de vérifier si les chaînes de référence sont l'un de ces traits linguistiques qui, s'il est contraint par le format IMRaD, permettent d'opposer les différentes parties « introduction », « méthodologie », « résultats », « discussion » et même « conclusion ».

De plus, puisque la plupart des entitées que les articles de recherche décrivent sont abstraites (c'est-à-dire que les articles discutent d'idées plutôt que de personnes ou d'objets), nous nous demanderons dans quelle mesure ces entitées peuvent être dites « avoir une référence », puis comment il est possible de classer de tels « référents abstraits », et sur quels critères (sémantiques, référentiels, ou directement liés à l'analyse des chaînes de référence qu'ils initient)

On pourra alors tenter de déterminer quelles sont les corrélations entre le comportement des chaînes de référence, le type de référent, et la partie IMRaD dans laquelle elles se trouvent, c'est-à-dire leur fonction rhétorique.

L'étude permettra à la fois de caractériser l'article de recherche de format IMRaD par l'étude des principales chaînes de référence qui y sont présentes ; et, du point de vue des chaînes de référence, elle permettra d'esquisser une « typologie » des chaînes dans les textes académiques et d'entamer une réflexion sur la corrélation entre ces différents types et les paramètres utilisés pour décrire les chaînes.

4 Organisation

Ce travail présentera d'abord la spécifité des référents abstraits et les problèmes qu'ils peuvent poser. Puis, dans trois chapitres plus méthodologiques, nous décrirons comment nous avons choisi le corpus, comment nous avons élaboré des outils d'annotation et d'analyses, et quel schéma d'annotation nous avons retenu. Enfin, nous présenterons deux études exploratoires, l'une pour laquelle nous avons annoté, pour chaque texte, un ensemble pré-sélectionné de référents saillants, l'autre pour laquelle nous avons annoté systématiquement toutes les « chaînes de paragraphe ».

5 Le projet Democrat

Nous terminerons cette introduction en évoquant le projet Democrat (« DEscription et MOdélisation des Chaînes de Référence : outils pour l'Annotation de corpus et le Traitement automatique »), auquel nous participons.

Il s'agit d'un projet financé par l'Agence Nationale de la Recherche (ANR) sur quatre années à compter de janvier 2016. Trois laboratoires y participent : le LaTTICe à Paris, LiLPa à Strasbourg, et ICAR à Lyon.

Le projet « vise à développer les recherches sur la langue et la structuration textuelle du français *via* l'analyse détaillée et contrastive des chaînes de référence (instanciations successives

d'une même entité) dans un corpus diachronique de textes écrits entre le 9ème et le 21ème siècle, avec des genres textuels variés[2] ».

Pour ce faire, le projet cherche à la fois à modéliser les chaînes sur le plan linguistique, et à annoter un grand corpus qui servira à l'apprentissage d'un système de détection automatique des chaînes de référence.

Même si nous participons à Democrat, nous avons défini notre sujet (octobre 2015) avant le début du projet (mars 2016), aussi ce travail se situe-t-il un peu à la marge, notamment en ce qui concerne le corpus, qui, pour des raisons techniques de droits d'auteur, ne pourra faire partie du corpus de Democrat. Par ailleurs, le schéma d'annotation du projet n'était pas encore arrêté lorsque nous avons annoté notre corpus : nous avons donc fait parfois des choix d'annotation différents de ceux de Democrat.

Nous espérons cependant que ce travail pourra contribuer à la réflexion sur l'annotation et l'analyse des chaînes de référence au sein du projet Democrat.

[2] http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2[CODE]=ANR-15-CE38-0008 , consulté le 30 mai 2016.

Chapitre 1

Le problème des référents abstraits

1 Introduction

Les études sur les chaînes de référence portent presque toutes sur des référents qui sont des humains ou des entités nommées (par exemple Chastain, 1975, Corblin, 1995, Schnedecker, 2005, Landragin, 2011, Schnedecker et Longo, 2012, Glikman, Guillot-Barbance et Obry, 2014) ou des objets physiques (par exemple Schnedecker, 2014).

Seules Longo et Todirascu (2014) étudient les chaînes de référents plus abstraits dans un corpus de textes juridiques, sans, cependant, s'interroger sur les problèmes de référence que posent de tels référents.

Or si une notion intuitive de la référence peut suffir pour l'annotation des chaînes de référents humains ou physiques, l'étude des chaînes de référents abstraits doit commencer par une réflexion plus globale sur la référence. Il convient d'abord de se demander si les noms abstraits réfèrent, et, si c'est le cas, de quelle nature est leur référent. En effet, si une chaîne de référence est définie comme la suite des expressions qui renvoient au même référent, encore faut-il savoir ce que veut dire « le même référent ». Cela ne pose que peu de problèmes avec des référents humains ou physiques (avec les réserves qu'observent, par exemple, Charolles et Schnedecker (1993) au sujet des « référents évolutifs[3] »), mais se révèle fort complexe pour les référents abstraits. Et encore faudrait-il montrer qu'on peut parler de « référents abstraits ».

Prenons un exemple qu'on pourrait rencontrer dans un article scientifique :

(4) Le taux de natalité n'est pas le même en France et en Allemagne. En effet, il est de 12 ‰ en France, alors qu'il est de 8 ‰ en Allemagne.

[3] Par exemple "un poulet actif et bien gras" qui, au cours de la même chaîne de référence, se transforme en poulet rôti.

Intuitivement, on aurait tendance à considérer *le taux de natalité* comme le premier maillon d'une chaîne qui continuerait par *il* et *il*.

Mais qu'en est-il dans l'exemple suivant, à peine modifié :

(5) Le taux de natalité n'est pas le même en France et en Allemagne. En effet, le taux français est de 12 ‰, alors que le taux allemand est de 8 ‰.

Peut-on vraiment dire que *le taux français* et le *le taux allemand* font partie de la chaîne initiée par *le taux de natalité* ? Le taux serait-il à la fois français et allemand ? Bien sûr, il s'agit de la valeur du taux, et non du taux lui-même (dont la formule mathématique ne varie pas en fonction du pays). Mais cela ne fait que créer un nouveau problème : faut-il initier une chaîne à chaque nouvelle valeur ? Dans ce cas, *le taux de natalité* n'initierait jamais aucune chaîne (puisque sa valeur change à chaque instant), ce qui est contre-intuitif par rapport à l'exemple (4).

Même avec un terme plus simple, comme *repas*, l'annotation est difficile :

(6) Ce week-end, j'ai mangé deux fois le même repas. Il était composé de frites. (...) Cependant, j'ai dû prendre le repas de samedi dans la précipitation, parce que j'avais une réunion après (...), alors que celui de dimanche était plus tranquille. (...)

Manifestement, le pronom *il* reprend le repas (ou plutôt deux repas), qu'il envisage sous l'angle du menu. Mais que faire des expressions *le repas de samedi* et *le repas de dimanche*, qui considèrent les deux repas sous l'angle du temps ? Reprennent-elles *le même repas* et *il* ? Initient-elles d'autres chaînes. Doit-on alors admettre qu'on a trois chaînes (*le même repas* et *il* ; *le repas de samedi...* ; *le repas de dimanche...*), pour seulement deux repas ?

Observons encore un exemple :

(7) Voici comment s'est déroulé le suicide de Werther : À minuit, il s'est tiré une balle dans la tête, mais sa mort n'est survenue qu'à onze heures le lendemain matin.

Doit-on considérer le *suicide* de Werther et sa *mort* comme le même événement ? Ou bien doit-on les séparer, en considérant que le suicide commence avec le coup de feu ?

De même, dans un article sur les pratiques alimentaires (texte T1), doit-on considérer comme différents la *consommation alimentaire*, *la consommation de protéines*, *la consommation de protéines animales*, *la consommation de viande* ? Ou bien toutes ces expressions (ou seulement certaines) doivent-elles être considérées comme renvoyant au même référent ?

La consommation de viande du groupe X le 1er février a-t-elle le même référent que *la consommation de viande du groupe X le 2 février* ? Et alors *la consommation de viande du groupe X* tout court ?

La mort de César est-elle le même événement que son assassinat ? (*A priori*, pas pour Davidson.)

Mais les événements ne peuvent-ils pas aussi s'exprimer par des verbes ? *La mort de César en 44 avant notre ère* ne représente-t-il pas le même événement que *César est mort en 44 avant notre ère* ? Faudrait-il seulement annoter la première expression, alors que les deux expressions reprennent le même événement ? Et que penser de *l'amour de Paul et Marie* et *Paul aime Marie* ?

Il conviendrait aussi de s'interroger sur la spécificité, la non-spécificité et la généricité de référents abstraits. Qu'est-ce qu'un référent abstrait non-spécifique ? *La consommation de viande du groupe X* est-elle une référence spécifique, non-spécifique ou générique ?

Il n'y a pas lieu de s'interroger ici sur l'existence ontologique de ces éventuels référents (puisque c'est le domaine de la métaphysique), mais on ne peut pas faire l'économie d'une réflexion sur l'existence (au moins dans la « métaphysique linguistique », comme dirait Asher) des entités auxquelles renvoient ces différentes expressions si on veut pouvoir annoter des chaînes de référents abstraits.

2 Le traitement traditionnel de la référence

2.1 Définition et expression

2.1.1 Définition

Huang (2014) définit la référence de deux manières différentes. Il commence d'abord par une définition *sémantique*, initiée par Frege : la référence est la relation entre une expression linguistique (l'expression référentielle) et un objet du monde externe (ou d'une représentation mentale). Il faut donc distinguer deux pôles : l'expression référentielle et le référent.

Mais il y a aussi une définition *pragmatique*, initiée par Strawson : la référence est l'acte par lequel le locuteur désigne un objet du monde extérieur (ou d'une représentation mentale) au moyen d'une expression linguistique (l'expression référentielle). Il faut donc ici distinguer trois pôles : l'expression référentielle, le référent mais aussi le locuteur.

Alors que Kleiber (1981) considère la référence plutôt dans le premier sens, Charolles (2002), par exemple, l'envisage plutôt dans le deuxième. Pour notre part, nous en resterons, dans un premier temps, à la définition sémantique, parce qu'elle est plus maniable, mais nous serons finalement contraint d'introduire le choix pragmatique du locuteur pour pouvoir individuer les différents référents.

Les expressions référentielles traditionnelles sont (pour l'anglais, mais aussi pour le français) :

- le nom propre et les descriptions définies,
- les SN démonstratifs,
- les SN indéfinis (y compris les pronoms indéfinis comme *quelqu'un*).

Cruse (2000, p. 313) y rajoute des déictiques, comme

- le pronom personnel (notamment de première et deuxième personne),
- certains adverbes de lieu (*ici, là, ...*),
- certains adverbes de temps (*maintenant, hier, ...*).

Nous dirons dans ce qui suit quelques mots sur les indéfinis, puisque ce sont eux qui semblent poser le plus de problèmes.

2.2 Les indéfinis : entre référence spécifique et non-spécifique

2.2.1 En philosophie analytique

En logique traditionnelle (Corblin, 1997, p. 9), les expressions référentielles, qui désignent (c'est-à-dire avec lesquelles on peut pointer un objet), s'opposent aux expressions quantifiées, qui ne désignent pas. Or toutes les expressions indéfinies sont considérées comme quantifiées, et donc aucune ne réfère.

Heim (1982) se demande si les indéfinis peuvent néanmoins référer. Elle retrace l'histoire des différentes positions que les philosophes ont pu tenir.

Ainsi, Russel considère que la phrase :

(8) Un chien est entré.

est toujours existentielle (il y a au moins un chien qui est entré), et ne désigne (c'est-à-dire ne réfère) jamais un chien en particulier.

Néanmoins, cette analyse est mise en cause par certaines reprises anaphoriques, comme dans :

(9) Un chien est entré. Il est couché sous la table.

Dans ce cas, *il* doit bien référer à la même entité que *un chien*, et donc *un chien* doit bien être une expression référentielle.

Trois solutions, poursuit Heim, peuvent être proposées dans le cadre de la théorie de Russel.

1. Pour Geach, il s'agit d'une variable liée, mais il faut admettre alors que la portée des quantifieurs dépassent le cadre de la phrase, ce qui pose d'autres problèmes : les propositions isolées n'ont plus de valeur de vérité.
2. Deux solutions sont dérivées de la philosophie de Grice :
   (a) Pour Kripke, il y a deux références :
      - la référence sémantique (*semantic reference*), définie par les règles de la langue. C'est la référence dont parle Russel,
      - la référence du locuteur (*speaker's reference*), définie par l'intention du locuteur : le référent est alors « ce dont le locuteur veut parler, ce qu'il a en tête » (p. 18)
   (b) Pour Lewis, le pronom réfère à l'objet dont la saillance est maximale. Ainsi, l'expression indéfinie (quantifiée et donc non référentielle) rend un objet saillant, de sorte que le pronom peut alors y référer.
3. Pour Evans (qui reconnaît et les variables liées de Geach et la saillance de Lewis), le pronom considéré est un *E-type pronoun* . Ces pronoms ont pour antécédent une expression quantifiée, sans y être lié. Au contraire, ils sont équivalents à une expression définie du type « le chien qui est venu ».

2.2.2 L'hypothèse de l'ambiguïté

D'autres philosophes, continue Heim, comme Strawson et surtout Chastain (1975), s'éloignent du cadre de Russel. Pour eux, les expressions indéfinies peuvent être ambiguës. Ainsi, dans la phrase (9), l'indéfini a bien une référence. Dans d'autres cas, cependant, comme dans la

négation ou lorsque des déterminants indéfinis comme *quelques* sont utilisés, les expressions indéfinies ne sont pas référentielles.

C'est cette « hypothèse » (comme la nomme Heim) qu'on retrouve généralement chez les linguistes (Kleiber, 1981, Cruse, 2000, Huang, 2014). Ainsi, pour Huang (2014) [4] :

(10) Hans veut aller dans une université américaine prestigieuse.

peut être :

- spécifique (Hans sait dans quelle université il veut aller),
- non-spécifique (Hans ne le sait pas, il n'a qu'une liste d'universités en tête). D'après Cruse (2000, p. 308), cette lecture n'est possible que dans certains contextes introduits par des verbes tels que *vouloir, devoir*.

Cruse (2000, p. 309) signale que la différence est codée grammaticalement dans certaines langues. Les exemples que donnent Cruse sont français :

(11) a. Marie cherche un homme qui *peut* lui faire l'amour douze fois par jour.
b. Marie cherche un homme qui *puisse* lui faire l'amour douze fois par jour.

Dans la première phrase, Marie sait qui elle cherche, dans la seconde, elle est seulement « *overly optimistic* ».

On considère généralement que seule la lecture spécifique donne lieu à une expression référentielle, ce qui est problématique pour l'annotation de certaine chaîne de référence. Par exemple, lorsqu'il est demandé dans une fiche de bricolage de « prendre un marteau », la référence à « un marteau » n'est pas spécifique (n'importe quel marteau fait l'affaire, et le marteau finalement utilisé sera différent d'un bricoleur à l'autre). Doit-on alors l'intégrer dans une chaîne *de référence* ?

(12) Prenez *un marteau* mais faites attention de ne pas *le* laisser tomber sur votre pied.

Les expressions indéfinies à référence non-spécifique sont en effet des expressions quantifiées (*trois tomates*), qui ne sont pas considérées, dans la tradition, comme référentielles (Corblin, 1987, p. 11).

Une autre approche possible est celle de la notion de « référent du discours ».

2.2.3 Le référent du discours

Karttunen (1976, p. 366) a introduit la notion de « référent du discours » pour « noter les référents mentionnés dans le discours, que les entités en question appartiennent ou non au monde réel » (Corblin, 1995, p. 17). Cette notion[5] permet donc de se passer de la question ontologique, *i.e.* de se demander quels sont les référents qui existent véritablement dans le monde.

[4] L'exemple est *Je veux épouser une tahitienne* chez Kleiber (1981) et *Mary veut épouser un banquier norvégien* chez Cruse (2000).

[5] Qui a aussi sa contrepartie en psycholinguistique : Langacker, 2008, p. 270.

Cependant, la question ontologique, une fois rapportée dans la « métaphysique linguistique », comme dirait Asher (1993), n'est pas inutile. En effet, Karttunen (1976, p. 366) propose qu'un indéfini ne puisse être dit n'introduire un référent du discours seulement si cela se justifie par la reprise anaphorique subséquente de ce référent. Il y a donc une implication sur l'annotation des chaînes de référence, notamment si l'on souhaite annoter les expressions référentielles hors chaîne pour pouvoir comparer les expressions qui forment une chaîne et celles qui n'en forment pas.

Karttunen (1976) s'interroge notamment sur les expressions indéfinies non-spécifiques (qui, en termes philosophiques, sont des expressions quantifiées et donc non-référentielles), dont toutes ne peuvent pas être reprises par une expression anaphorique. Par exemple (en prenant les exemples strasbourgeois de Kleiber (1981, pp. 146 *sq.*)) :

(13) a. Je veux épouser une tahitienne (= spécifique). Elle s'appelle Maeva.
b. Je veux épouser une tahitienne (= non-spécifique). * Elle est très belle.

On voit ici que l'interprétation non-spécifique empêche la reprise nominale. Karttunen dirait que *une tahitienne* dans (13a) n'introduit pas un référent du discours. Ce qui a une répercussion sur l'annotation : devrait-on l'annoter ?

En fait, ce phénomème se rencontre lorsque l'on modalise le discours (*il faut, désirer, vouloir, espérer, proposer*, etc.). Mais, continue Karttunen (1976, p. 374), on peut faire référence à un indéfini non-spécifique du moment qu'on reste dans la « portée » de la modalité :

(13) c. Je veux épouser une tahitienne (= non-spécifique). Elle devra être très belle.

On peut traiter ces cas en faisant appel à la notion de « variable liée », issue de la logique du premier ordre. On parle de « variable liée » lorsque le pronom dépend d'une expression quantifiée (Huang, 2014, p. 235). Cependant, on peut peut-être aussi en appeler au concept de « monde possible » (ce qui se traduit en termes psycholinguistiques par « espace mental » selon Croft et Cruse (2004, p. 33)). Mais si Kleiber (1997) ne semble pas être en faveur de cette notion, elle semble pourtant pratique pour décrire le phénomène pointée par Karttunen : un référent de discours ne dure que le temps que dure le monde dans lequel il a été introduit. Ainsi, je peux introduire la tahitienne (non-spécifique) que je veux épouser dans un monde possible, et donc y faire référence tant que ce monde possible dure (c'est-à-dire, d'un point de vue cognitif, tant que je le maintiens en mémoire, ou, d'un point de vue logique, tant qu'on se trouve dans sa portée). C'est d'ailleurs ce qui permet de traiter les chaînes de référence dans les recettes de cuisine ou les fiches de bricolage, dont les référents sont pour la plupart des indéfinis non-spécifiques. On peut y référer parce que ces textes établissent des mondes possibles/espaces mentaux/portées qui durent sur l'ensemble du texte.

Cela nous semble un élément dont il faut tenir compte pour l'étude et l'annotation des chaînes de référence : la portée d'une chaîne peut s'expliquer en partie par la modalité ou le type du verbe où l'expression indéfinie apparaît.

3 La référence des noms abstraits et des prédicats

Dans un premier temps, nous ne traiterons que des noms prédicatifs. Nous introduirons les autres noms abstraits plus tard dans l'analyse.

3.1 Qu'est-ce qu'un prédicat ?

On oppose généralement expression référentielle et prédicat (Kleiber, 1981, Charolles, 2002). Néanmoins, la sémantique formelle nous semble proposer une définition plus opératoire pour notre propos. Ainsi, Corblin (2013, pp. 100 *sqq.*) définit le prédicat comme « une expression qui produit une proposition si elle est associée à une (ou plusieurs...) constante individuelle ». Les constantes sont des éléments de l'univers de référence, qui remplacent des places laissées vides dans le prédicat. Par exemple, dans

(14) x mange y

x et y sont des variables. On peut les actualiser par des constantes, comme « chat » et « souris » :

(15) Le chat mange la souris.

Cette définition vient de la définition de Frege (1971b [1891]), qui compare le prédicat à une fonction mathématique. Une fonction est dite « incomplète, ayant besoin d'autre chose, ou encore insaturée » (Frege, 1971b [1891], p. 84).

Chez Frege, les prédicats peuvent renvoyer des objets, tout comme les fonctions mathématiques. Ainsi, le prédicat « capitale de x » peut renvoyer « Berlin » si x est « empire allemand » (à l'époque de Frege). Ce qui permet d'imbriquer les prédicats les uns dans les autres.

Pour que cette imbrication puisse avoir lieu, il faut que le prédicat, qui a souvent une forme verbale ou adjectivale, prenne une forme nominale (Kleiber (1981, p. 83) explique ainsi le passage de *blanc* à *blancheur*). Ce sont des prédicats nominaux, selon la terminologie de Gross (2012).

3.2 Les prédicats référent-ils ?

3.2.1 La réponse de Kleiber

Kleiber (1981) commencent par dénier aux noms syncatégorématiques, aux adjectifs et aux verbes toute référence. Ainsi, il affirme (p. 58) que seuls les noms catégorématiques réfèrent, et que « il est naturel de la [*i.e.* la référence] refuser aux verbes et aux adjectifs » (p. 67).

Dans le même temps, pourtant, il semble reconnaître aux noms syncatégorématiques, aux adjectifs et aux verbes une référence, mais d'une autre nature : « verbes et adjectifs ne pourront jamais être dits référer comme les noms catégorématiques, parce qu'ils renvoient à "quelque chose de plus réel" » (p. 66).

En fait, on pourrait dire que, pour Kleiber, la référence se confond avec une présupposition d'existence : « Notre hypothèse sera que les items lexicaux *réfèrent* parce qu'ils présupposent l'existence d'un référent conceptuel » (p. 15). Cette conception de la référence lui permet de classer les substantifs en trois catégories (pp. 66–67) :

- les catégorématiques individuants (comptables),
- les catégorématiques globalisants (massifs),
- les syncatégorématiques, qui sont l'équivalent nominal des adjectifs et des verbes.

La référence des deux premières classes ne pose pas de problèmes, parce qu'ils présupposent l'existence de leur référent. Les syncatégorématiques, les adjectifs et les verbes supposent qu'on examine plus attentivement la notion de *référence*. On découvre alors qu'il y a « deux sens de *référer* » (p. 111) :

- le sens propositionnel peut être paraphrasé par « occuper la position référentielle » et peut être employé dans une phrase comme « le locuteur se sert de cette expression pour référer à… » (p. 112). Par exemple, dans « Je mange cette pomme », on dira que *pomme* réfère à l'objet que j'ai (ou dont une partie est) actuellement dans ma bouche, et c'est parce que je peux désigner l'objet dont je parle qu'il y a référence (au sens propositionnel),
- le sens non-propositionnel, lui, est le sens qu'a *référer* dans une phrase comme « cette expression réfère, parce qu'elle présuppose l'existence de… » (p. 112). Ce type de référence est, ajoute Kleiber, l'équivalent de la référence virtuelle de Milner : le prédicat (puisqu'il s'agit d'un prédicat) est alors considéré hors de toute proposition (tout comme les noms catégorématiques ont une référence virtuelle hors de toute détermination). Par exemple, le prédicat « *être ivre* présuppose l'existence d'un référent *ivresse* » (p. 128), et c'est parce qu'il y a « cette présupposition d'existence » (p. 15) qu'il y a référence (au sens non-propositionnel).

Ces deux sens de *référer* sont construits à partir de l'« asymétrie référentielle » (p. 97), concept que Kleiber reprend de Strawson. Ainsi, lorsque Kleiber dit que les prédicats ne peuvent pas référer comme les noms catégorématiques, c'est qu'il y a une différence fonctionnelle : dans la proposition basique de type sujet-prédicat (*Paul chante*), le sujet (logique) est « plus » référentiel que le prédicat.

La réponse de Kleiber à la question (qu'il se pose explicitement (pp. 127 *sqq.*)) « Les prédicats réfèrent-ils ? » trouve une réponse à deux niveaux :

- « oui » au sens de la référence non-propositionnelle,
- « non » au sens de la référence propositionnelle.

Le problème, c'est que la référence non-propositionnelle (tout comme la référence virtuelle de Milner) n'est d'aucune utilité pour l'annotation d'un texte, où l'on ne rencontre *que* des propositions. Donc, pour Kleiber, la réponse à la question « Les prédicats réfèrent-ils *dans un texte* ? » est « non ». Ce qui signifie qu'il ne faudrait considérer comme pouvant faire partie d'une chaîne *de référence* que les noms catégorématiques (comptables ou non), puisque, d'après ce que nous comprenons, les noms syncatégorématiques sont à assimiler aux adjectifs et aux verbes (*cf.* p. 58, où Kleiber « range[] les substantifs syncatégorématiques, parce qu'on les tient pour des nominaux dérivés, avec les adjectifs et les verbes »). Ce sont alors des prédicats.

3.2.2 La réification

On a pourtant l'intuition que

(16) Marie aime Pierre.

et

(17) L'amour de Marie pour Pierre.

renvoie à la même entité. Ou, comme le dit Davidson (2004 [1966], p. 737), *"there are such* things *as actions"* , actions que des phrases peuvent décrire de différentes façons. *"'Jones did it with a knife.' 'Please tell me more about it.' The 'it' here doesn't refer to Jones or the knife, but to what Jones did."* Ailleurs, Davidson (2006 [1969], p. 90) parle de *"singular terms referring to events"* .

Il s'agit en fait d'une « réification », que Seuren (1998, pp. 385–386), du côté de la sémantique formelle, décrit ainsi :

> *A reification comes about when a mental construct of some complexity is captured, more or less by way of definition, in a single nominal expression... They are problematic, obviously, because if they are definite descriptions they are in search of a reference object, and no sound ontology could possibly provide entities corresponding to such expressions. [They are] an abstraction, based on some more or less precise calculus of values on selected parameters. Adequate interpretation of sentences containing reifications requires knowledge of the conditions under which reifying expressions are to be applied to the objects of the world.*

Et plus loin : *"Yet they are treated linguistically as if they were entities, since they are referred to by referring terms."*

Cette notion se retrouve aussi chez Langacker (2008) (pp. 105–108 et p. 120), bien qu'il maintienne une différence fondamentale entre les prédicats (« verbes ») et les objets (« noms »).

On peut donc admettre que les prédicats, même s'ils ne réfèrent pas dans le sens ontologique du terme, renvoie à une entité unique (une réification).

3.3 Faut-il annoter les verbes ?

Il existe une relation manifeste entre les verbes et certains noms : « un grand nombre de substantifs du français peuvent se classer et s'analyser comme les verbes et les adjectifs, à savoir comme des prédicats assortis d'un certain nombre d'arguments (les sujets et les compléments) » (Giry-Schneider, 1987, p. 1). Gross (2012) parle de « prédicats nominaux » ou de « noms prédicatifs », qu'il définit comme des substantifs qui ont une structure argumentale.

De fait, les prédicats ont différentes réalisations morphologiques (Gross, 2012, p. 36) : verbe, adjectif ou substantif (parfois les trois, parfois deux seulement, parfois seulement un nom). Gross et Vivès (2001, p. 39) précisent que « les prédicats appartiennent à plusieurs catégories—verbes, substantifs, adjectifs prédicatifs essentiellement—mais l'analyse de base est fondamentalement la même ». Par exemple :

(18) a. Pierre admire ce tableau
b. Pierre a de l'admiration pour ce tableau
c. Pierre est admiratif devant ce tableau

(*Être* et *avoir* sont des verbes « support ».)

Nous reviendrons plus tard sur les noms prédicatifs de Gross et nous choisirons une définition un peu différente, mais il suffit de noter pour l'heure que le prédicat a plusieurs formes.

On se limite généralement à l'annotation des noms (Longo et Todirascu, 2014, p. 95), mais si un prédicat peut s'exprimer aussi bien sous la forme d'un nom, d'un adjectif ou d'un verbe, pourquoi faudrait-il se limiter aux seuls noms ? L'analyse des chaînes de référence par Chastain offre un exemple éclairant sur l'importance de l'annotation de *tous* les prédicats.

La notion de chaîne de référence a été introduite par Chastain (1975, p. 204) (*cf.* Schnedecker et Landragin, 2014, Corblin, 1995, p. 151), au cours de la présentation d'une théorie de la référence. Il ne considère que les « termes singuliers », c'est-à-dire les expressions référentielles, au sens philosophique du terme, c'est-à-dire les termes qui désignent (ou nomment) une entité du monde réel. Il n'est pas question, pour lui, d'annoter des expressions quantifiées ou des référents génériques ou fictionnels (d'après ce que nous comprenons).

Chastain donne l'exemple suivant :

(19) At eleven o'clock that morning, an ARVN officer stood a young prisoner, bound and blinfolded, up against a wall. He asked the prisonner several questions, and when the prisoner failed to answer, beat him repeteatedly. An American observer who saw the beating reported that the officer « really worked him over ». After the beating , the prisonner was forced to remain standing against the wall for several hours.

qui contient les chaînes de référence[6] suivantes :

- *that morning*
- *an ARVN officer, he, the officer*
- *a young prisoner, the prisonner, the prisonner, him, him, the prisonner*
- *a wall, the wall*
- *an American observer who saw the beating*
- *the beating, the beating*

C'est la dernière chaîne qui nous intéresse. On remarque d'abord que si Chastain dit ne s'intéresser qu'aux « individus concrets » (p. 195), il annote *the beating* qui est clairement un « nom

[6]Chastain appelle ces chaînes des « chaînes anaphoriques » et les distingue des « chaînes de référence », qui s'étendent sur plusieurs contextes. Par exemple, le lecteur d'un journal peut lire un article sur le Dr. Michael DeBakey (Chastain, 1975, p. 212) et en parler plus tard dans une conversation avec sa femme. Il y aura un lien entre les deux expressions renvoyant à DeBakey, celle dans l'article de journal et celle dans la conversion. Ce lien sera « référentiel », et les expressions formeront une « chaîne référentielle ». Par contre, toutes les expressions à l'intérieur d'un même contexte (l'article *vs* la conversion) sont liées par une « chaîne anaphorique ». Cette distinction n'a pas lieu d'être en linguistique, mais n'oublions pas que Chastain était philosophe, et qu'il décrit ces notions dans un article qui expose une théorie de la référence (donc une théorie philosophique). Corblin (1995) fait une distinction entre « chaînes-A » (anaphoriques) et « chaînes-R » (référentielles), mais la distinction n'a plus rien à voir avec celle de Chastain, et est purement linguistique (une chaîne anaphorique contient des anaphores, une chaîne référentielle des coréférences, au sens linguistique des termes).

prédicatif » (ou un « nom syncatégorématique » dans la terminologie de Kleiber (1981)), et qui n'est en rien « concret ». C'est probablement son intuition qui a guidé Chastain à l'annoter, ce qui montre bien, pensons-nous, l'importance d'annoter les noms prédicatifs.

Cependant, ce qui nous intéresse surtout ici, c'est l'explication que donne Chastain de la dernière chaîne. Il remarque d'abord que toutes les chaînes commencent par un indéfini (puisque l'indéfini a un rôle introducteur), sauf la première (qui ne pose pas problème), et la dernière. Corblin (1995, p. 153) remarque que Chastain semble admettre (même s'il ne le fait explicitement) que « le premier terme d'une chaîne anaphorique est nécessairement une expression indéfinie ». Ce qui est confirmé par le fait, croyons-nous, que Chastain dit que les expressions définies peuvent seulement « continuer » des chaînes anaphoriques, et ne peuvent donc pas les initier.

Or la dernière chaîne pose problème puisqu'elle ne commence pas par un indéfini. Chastain postule donc la présence, « *after a deeper analysis* », d'une expression indéfinie *ad hoc* : « *a beating of the yound prisonner by the ARVN officer* ». Corblin se contente de remarquer le fait en citant Chastain, sans chercher d'explication.

Il nous semble qu'il y a cependant une meilleure explication. En prenant en compte, comme maillon de chaîne, des verbes, il n'y a plus besoin de postuler des expressions indéfinies qui ne sont pas là, puisque la chaîne contenant *the beating* serait initiée par le verbe *beat* . Les explications de Chastain seraient alors beaucoup plus convaincantes.

Un autre exemple, tiré du corpus utilisé par Longo et Todirascu (2014, p. 95), va dans le même sens :

(20) L'article 6... interdit-il aux autorités nationales compétentes de [retirer le permis de séjour d'un travailleur turc]$_i$, qui ne s'est rendu coupable d'aucun comportement frauduleux, avec effet rétroactif à la date à laquelle le motif auquel le droit national subordonnait l'octroi du permis de séjour a cessé d'exister, [ce retrait]$_i$ intervenant après l'expiration du délai d'un an visé à l'article 6, paragraphe 1, premier tiret, susvisé ?

Comment pourrait-on justifier l'emploi du démonstratif *ce retrait* (qui n'est pas ici déictique, mais anaphorique) si on ne tient pas compte de *retirer* ?

Ces exemples montrent, à notre avis, l'importance ou bien d'annoter *tous les prédicats*, qu'ils soient nominaux, verbaux, ou même adjectivaux[7], ou bien de n'en annoter *aucun*.

De même, Longo et Todirascu (2014, p. 95) « propos[ent] d'inclure dans les maillons des chaînes de référence des textes non-narratifs étudiés des verbes (conjugués ou à l'infinitif) référant à des entités abstraites dans le discours (actions, événements, faits) », car cela « permettrait de mettre en évidence des propriétés des référents non-humains ».

La prise en compte de tous les prédicats, y compris sous leur forme verbale, permettrait également d'étudier l'évolution de la structure argumentale. Par exemple, Condette, Marin et Merlo (2012) étudient la différence des structures argumentales entre les prédicats nominaux et les prédicats verbaux, et concluent que les prédicats nominaux ont généralement moins d'argu-

[7] Mais peut-être pas tous de la même manière, c'est-à-dire avec une propriété qui permettent de distinguer les verbes, les adjectifs et les noms.

ments explicites que les prédicats verbaux. On pourrait faire l'hypothèse que c'est parce ce que les « chaînes de prédicats » commencent par un verbe, plus à même de spécifier la totalité des arguments. C'est ce qu'il se passe d'ailleurs dans l'exemple de Chastain (*he beats him* est repris simplement par *the beating*).

3.4 L'individuation et l'identité des prédicats

Si on accepte, dans les chaînes référentielles, les prédicats, qu'ils soient sous forme de verbes, de noms ou d'adjectifs, encore faut-il savoir comment les individuer et les identifier (ce qui revient au même, comme le dit Quine (2013 [1960], p. 105)).

3.4.1 L'identité formelle

Une première approche consiste à ne considérer que l'identité formelle. C'est l'approche utilisée en TAL, pour des raisons pratiques, et c'est donc l'approche de Longo et Todirascu (2015), qui indiquent ne traiter que « les situations de coréférence directe… où les groupes nominaux coréférents possèdent la même tête nominale » (p. 134) (bien que leur exemple 8 fasse apparaître des têtes nominales différentes !). Cependant, comme le souligne Corblin (1995, p. 174), les chaînes de référence de la langue naturelle ne reposent pas sur l'identité formelle. Il faut donc tenir compte des différentes variations lexicales (au moins).

3.4.2 L'individuation des événements selon Davidson

Bien qu'il ne traite que des événements, la réflexion de Davidson pourrait être utile pour savoir comment individuer des concepts abstraits.

Davidson (2006 [1969], p. 90) se demande en effet sur quel critère fonder l'identité ou la différence entre deux événements. Il rappelle que deux événements sont toujours distincts (le même événement ne se répète en effet jamais), mais qu'on peut se demander si deux termes singuliers (par exemple « l'éruption du Vésuve en 1906 » et « le Vésuve est entré en éruption en 1906 ») peuvent référer tous deux aux mêmes événements, et comment le savoir.

Le philosophe soulève des points dont il faut tenir compte ici. Par exemple, *le coup de poignard reçu par César* et *la mort de César* correspondent-ils aux mêmes événements ? En effet, ils ne coïncident pas dans le temps. Qu'on songe au suicide de Werther : il se tire une balle de fusil dans la tête à minuit, mais meurt à onze heures le lendemain matin : les événements *le suicide*, *le tir*, *la mort de Werther* sont-ils tous l'expression d'un même évévement ? De même un empoisement est-il le même événement que le versement du poison dans une nourriture et/ou que la mort de la victime ? Comment décider ?

Après avoir étudiés plusieurs hypothèses, notamment celles qui consistent à dire que deux événements sont identiques s'ils ont lieu au même endroit, ou au même moment, ou les deux (c'est le choix de Lemmon), Davidson conclut que deux événements sont identiques s'ils ont exactements les mêmes causes et les mêmes effets.

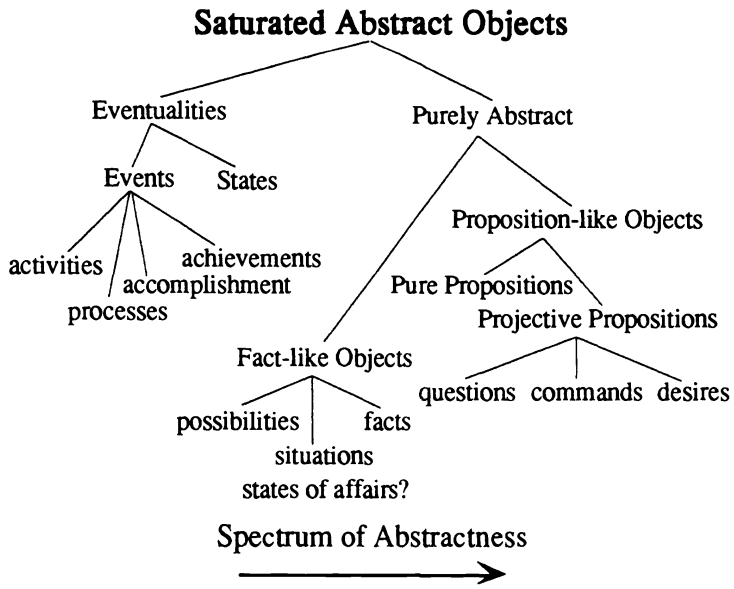

FIG. 1: Typologie des entités abstraites selon Asher, 1993, p. 57.

Cependant, cette définition n'est pas très opératoire, et n'aide guère à reconnaître, lors de l'annotation d'un texte, si deux noms abstraits ont le même référent (comment savoir, avec cette définition, si *le suicide*, *le tir*, *la mort de Werther* sont l'expression du même événement ?). Étudions donc l'ouvrage d'Asher (1993), dont le titre semble prometteur (*Reference to abstract objects in discourse*).

3.4.3 La référence des noms abstraits selon Asher

La figure 1 montre la typologie compliquée des entités abstraites qu'élabore Asher (1993). Le plus intéressant, dans ce diagramme, est sans doute la flèche qui indique une *différence de degrés* dans l'abstraction : tous les référents abstraits n'ont pas le même degré d'abstraction. Cela s'explique par la notion d'*immanence*, c'est-à-dire la capacité à avoir des propriétés physiques et à pouvoir interagir avec le monde physique. Par exemple, une entité qui a des propriétés spatiales et temporelles, mais aussi causales (c'est-à-dire la possibilité d'être une cause), sera considérée comme plus immanente au monde (*world immanent objects*) qu'une entité qui n'a qu'une propriété causale (*intermediate entities*), qui sera elle-même plus immanente qu'une entité purement abstraite (*purely abstract objects*).

Sans entrer dans toutes les ramifications de sa typologie (p. 57), Asher (1993) distingue (voir aussi Asher, 2000) :

- les objets concrets,
- les objets abstraits :
  - les événements et les états (qui ont des propriétés spatiales, temporelles et causales, *i.e.* ils peuvent être des causes),
  - les faits (qui n'ont pas de propriétés spatiales ni temporelles, mais qui ont des propriétés causales),
  - les propositions (qui n'ont aucune des ces trois propriétés).

Les objets concrets (*pomme, table*) ne posent pas de problèmes, et Asher ne s'y intéresse pas. Les événements (*l'assassinat de César*) sont plus complexes, mais Asher s'en remet à la distinction que Vendler avait faite en 1957. L'auteur s'attarde surtout sur la différence entre les faits et les propositions. Un fait peut, par exemple, être une cause (ou être le résultat d'une cause) :

(21) Le fait que John a eu mal à la tête l'a rendu grincheux. (Asher, 2000)
= Le mal de tête de John l'a rendu grincheux.

Au contraire, les propositions n'ont pas de pouvoir causal. « *How could they, given that these propositions do not depend for their existence on any realisation?* » (Asher, 2000, p. 128). Par exemple,

(22) [Il n'est pas vrai que Mary perdra la course]$_i$; Fred [le]$_i$ pense aussi. (Asher, 1993)

Ici, *Il n'est pas vrai que Mary perdra la course* n'a pas de propriétés spatio-temporelles, et n'a pas non plus de pouvoir causal. De plus, il peut être repris par un pronom (c'est l'anaphore conceptuelle que Asher (1993) développe dans la suite de son ouvrage).

Puisqu'on peut référer aux objets abstraits à l'aide d'expressions de divers types, comme des verbes ou des phrases, mais aussi des noms, doit-on les intégrer dans des chaînes de référence ? Nous pensons que oui, puisque, comme le dit Asher (1993, p. 7), on y fait constamment référence dans le discours. Or si on y fait référence, c'est qu'on peut les individuer. Mais c'est là tout le problème, continue Asher (1993, p. 7) : contrairement à la « vraie métaphysique », celle des langues naturelles n'offre aucun principe d'individuation. Seule l'interprétation sémantique, dépendante du contexte, permet d'identifier ces « entités du discours ».

Par exemple (Asher, 1993, pp. 7–8), on peut postuler deux événements au niveau du langage naturel, qui, au niveau métaphysique, n'en sont qu'un seul. C'est le cas bien connu (voir aussi, par exemple, Quine, 2013 [1960], p. 117) de l'ambiguïté « action-résultat » (selon le terme de Quine) :

(23) Marie passa le pas de la porte. Pierre faisait la vaisselle. Elle l'embrassa, puis s'enfonça dans son fauteuil. La journée avait été dure. (Asher, 1993)

Ici, l'événement *Pierre faisait la vaisselle* est décrit comme un « état » (au sens, dit Asher, que l'état ne contient ni son moment initial, ni son moment final). Mais le même événement (au niveau métaphysique) peut être décrit d'une autre façon, comme une action :

(24) Pierre fit la vaisselle. (Asher, 1993)

On décrit alors le même événement à la fois comme un état et une action.

Le problème, dit Asher (1993, p. 6), c'est qu'on ne passe pas sa vie cognitive à tenter d'individuer les actions, ni à tenter de déterminer (en dehors des cercles philosophiques) si deux événements sont identiques ou non. Il n'y aurait donc pas, pour Asher, de principe absolu qui permettrait de faire ces individuations ou ces identifications. Au contraire celles-ci sont faites en fonction du contexte. Dans certains cas, les événements sont individués selon leur description, dans d'autres cas, selon leur inscription spatio-temporelle. En ce qui concerne les autres objets abstraits, comme les propositions, Asher (1993, p. 6) propose de se fonder avant tout sur le but de la communication.

L'étude d'Asher (1993) se veut une étude des entités abstraites selon une sémantique formelle, à partir de l'analyse des expressions qui les dénotent et des phénomènes d'anaphore. Cependant, il s'inscrit dans le cadre de la *Discourse Representation Theory* (DRT, dont on trouve une introduction en français dans Corblin, 2002), qui dépasse le cadre de ce travail.

Ce que nous retenons de l'étude d'Asher, c'est surtout l'absence d'un principe absolu d'individuation. Au contraire, l'individuation et l'identification des référents abstraits se fait en utilisant telle ou telle propriété, et toujours en fonction d'une visée de communication.

3.4.4 L'anaphore des noms prédicatifs selon Gross

Gross (2006) examine les différentes anaphores de noms prédicatifs, ce qui pourrait nous aider à déterminer comment établir l'identité de deux concepts.

Nous rappelons que pour Gross, le sens d'un prédicat est déterminé par sa structure argumentale : un même prédicat peut avoir plusieurs sens (Gross dit « emplois ») s'il y a plusieurs schémas d'arguments (Gross et Vivès, 2001, p. 39).

Ce que Gross étudie, ce sont donc des anaphores résomptives, ou plutôt, comme il le dit (Gross, 2006, p. 369), des « anaphores argumentales » qui codent des informations syntaxiques.

Gross liste les différentes anaphores qui peuvent reprendre chacun des types de prédicats (qu'ils soient sous forme de nom, de verbe ou d'ajdectif) qu'il identifie : prédicats d'action, d'évévenement ou d'état. Par exemple les prédicats d'action (ex. : *assassiner*) peuvent être repris par un verbe (soit par répétition du verbe, soit par un hyperonyme), par un substantif (de même racine que le verbe (ce que Gross identifie comme une anaphore fidèle), ou bien un hyperonyme, ou un synonyme plus ou moins subjectif). Les prédicats d'événement (ex. : *séisme*) peuvent être pronominalisés avec *le, y*, remplacés par *comme ça, de la sorte* ou bien par un substantif comme *situation, condition*, etc. Enfin, les prédicats d'état (ex. : *être gentil*) peuvent être pronominalisés par *le*, ou bien repris par un substantifs, soit un hyperonyme, soit un hyponyme.

Cette approche confirme l'idée de traiter les verbes et les noms comme maillons potentiels d'une même chaîne de référence. Elle nous apprend aussi qu'il est difficile de juger si deux référents renvoient au même référent, tant les expressions linguistiques qui peuvent faire anaphore sont diverses.

Cependant, si elle est opératoire pour des anaphores isolées, cette approche ne s'intéresse pas à ce qu'est la « référence » ou son suivi. Elle n'est donc pas utile, par exemple, pour annoter les exemples proposés en introduction.

3.4.5 La prise en compte de la structure argumentale

Les noms prédicatifs, qui nous intéressent ici, ont donc une structure argumentale. Lorsqu'ils sont employés dans un énoncé, ils sont saturés par des arguments. Ces arguments pourraient donc être un moyen de les différencier et de les identifier. Par exemple, « l'assassinat de César » n'est pas « l'assassinat de Cicéron ».

Considérons qu'un prédicat a une liste d'arguments ainsi que des propriétés spatio-temporelles. Certains arguments seront plus importants que d'autres pour individuer un référent. Par exemple, s'il s'agit de comparer l'assassinat de César et celui de Cicéron, la victime de l'assassinat devra bien sûr être prise en compte pour individuer les deux assassinats. Mais dans une étude qui évalue le taux de meurtres à l'aune de telle ou telle politique, cet argument n'aura que peu d'intérêt. De même, s'il s'agit de comparer les taux de natalité français et allemand, la propriété spatiale aura une grande importance, mais pas la propriété temporelle. Ce serait le contraire, par contre, dans une étude qui étudierait l'évolution du taux de natalité en France depuis la Révolution.

Cette idée s'inpire de la sémantique lexicale de Cruse (1986, p. 53 et 2000, pp. 57–58), notamment des notions de *foregrounding* et *backgrounding* .

Cependant, comme bien d'autres idées présentées ici, ce principe risque de ne pas être très opératoire pour l'annotation d'un texte entier. On peut bien faire la différence entre les concepts sur un petit extrait, ou bien à un niveau général, mais il est difficile de maintenir cette différence sur tout un texte.

3.4.6 Des tests linguistiques

Un certains nombre de tests linguistiques peuvent être introduits pour tenter d'individuer les référents abstraits.

Il est *a priori* possible de suivre un même référent par des phénomènes d'anaphore. On pourrait, par exemple, utiliser le pronom *il* pour vérifier « que l'on continue à parler de la même chose » (Charolles et Schnedecker, 1993, p. 121).

Mais ce principe est d'application délicate avec les référents abstraits. En effet, le texte suivant semble indiquer qu'on a trois référents distincts, *le taux de natalité*, *le taux français*, *le taux allemand*, chacun pouvant initier une chaîne :

(25) [Le taux de natalité]$_i$ n'est pas le même en France et en Allemagne. En effet, [le taux français]$_j$ est de 12 ‰, alors que [le taux allemand]$_k$ est de 8 ‰.

Pourtant, si on exprime la même phrase avec des pronoms, on a plutôt l'impression qu'il n'y a qu'un seul référent, et qu'une seule chaîne :

(26) [Le taux de natalité]$_i$ n'est pas le même en France et en Allemagne. En effet, [il]$_i$ est de 12 ‰ en France, alors qu'[il]$_i$ est de 8 ‰ en Allemagne.

Cela indique que le test du pronom n'est pas fiable. Il permet cependant de décider quand une chaîne s'arrête. On ne peut pas continuer l'exemple (26) avec un pronom, ni même avec un démonstratif, qui réfèrerait à la chaîne i :

(27) [Le taux de natalité]$_i$ n'est pas le même en France et en Allemagne. En effet, [le taux français]$_j$ est de 12 ‰, alors que [le taux allemand]$_k$ est de 8 ‰. *Il/*Ce taux permet de comprendre les différences dans les politiques familiales des deux pays.

alors qu'on peut continuer la chaîne de l'exemple (25) :

(28) [Le taux de natalité]$_i$ n'est pas le même en France et en Allemagne. En effet, [il]$_i$ est de 12 ‰ en France, alors qu'[il]$_i$ est de 8 ‰ en Allemagne. [Il]$_i$ permet de comprendre les différences dans les politiques familiales des deux pays.

Une fois que la chaîne initiale est divisée, donc, il n'est plus possible d'y revenir. On peut cependant utiliser un hyperonyme pour revenir à la chaîne i :

(29) [Le taux de natalité]$_i$ n'est pas le même en France et en Allemagne. En effet, [le taux français]$_j$ est de 12 ‰, alors que [le taux allemand]$_k$ est de 8 ‰. [Cet indicateur]$_i$ permet de comprendre les différences dans les politiques familiales des deux pays.

Il faut toutefois noter que c'est problablement le prédicat (*permet de comprendre la différences dans les politiques familiales des deux pays*) qui permet de calculer la référence de *cet indicateur*. Sans ce prédicat, en effet, on ne saurait pas si le référent est i ou k.

Cruse (2000, p. 131) indique que sous *le taux de fécondité*, on pourrait trouver une référence à *la valeur* du taux (*covert reference*). Cela est possible, mais cela ne résout pas le problème d'un point de vue linguistique.

Le jeu des pronominalisations permet de repérer les différentes « facettes » d'un référent (en applicant la notion de « facettes » telle que décrite par Cruse (2000, pp. 114 *sq.*) aux référents abstraits). L'exemple suivant correspond parfaitement à ce que dit Asher (1993) : un événement est parfois considéré selon sa description, parfois selon son inscription spatio-temporelle :

(30) Aujourd'hui, j'ai mangé [le même repas qu'hier]$_i$. [Il/Ce repas]$_i$ était composé de frites. Cependant, j'ai dû prendre [le repas d'aujourd'hui]$_j$ dans la précipitation, parce que j'avais une réunion après, alors que [celui d'hier]$_k$ était plus tranquille. *[Il]$_{??}$/ ?[Ce repas]$_{k?}$/Quoi qu'il en soit [ce repas]$_i$ était vraiment délicieux.

On peut se demander ici, en effet, ce qui fait le repas : le menu ou la date ? Les deux, puisque (30) semble d'abord initier une chaîne avec *le même repas qu'hier*, chaîne fondée sur l'aspect descriptif du référent, alors que *le repas d'aujourd'hui* et *celui d'hier* initie de nouvelles chaînes en se fondant sur l'aspect temporel des événements. On remarquera d'ailleurs que *celui d'hier* n'est pas ici une description définie, mais un démonstratif qui illustre bien la fonction recatégorisante, ou reclassifiante, du démonstratif (voir par exemple Charolles, 2002, p. 120 et de Mulder, 1998). Il y a recatégorisation, car le repas n'est plus vu du point de vue du menu, mais du point de vue du temps.

L'exemple (30) présente un autre problème. Non seulement *le même repas qu'hier* initie une chaîne qui correspond en fait à deux événements, donc à deux référents, mais en plus on obtient à la fin trois chaînes... pour les mêmes deux événements. Il faut donc admettre que le nombre de chaînes ne correspond pas au nombre de référents. Ce ne sont donc plus à proprement parler des « chaînes de référence ». Le même problème apparaît dans un exemple donné par Huang (2014, p. 255) :

(31) [[Platon]$_i$]$_j$ est sur l'étagère du haut. [C]$_j$'est un grand philosophe. J'aime beaucoup [le]$_i$ lire, même si je ne comprends absolument rien à [sa]$_j$ philosophie.

La seule solution que nous puissions voir, c'est de créer deux chaînes qui se partagent un premier maillon : *Platon*. Mais là encore, il faut se demander si on peut encore parler de « chaîne de référence », puisque le premier maillon appartient à deux chaînes totalement différentes.

L'exemple (30) montre aussi qu'il est difficile de revenir à la chaîne initiale une fois qu'elle a été divisée. Cela était possible dans l'exemple (29). Mais ici, il ne semble pas possible de revenir en arrière, même avec un hyperonyme. Seul l'« annulation » des phrases dans lesquelles la division a été faite, avec le connecteur *quoi qu'il en soit* (qui permet de « revenir en arrière » dans le discours, en rejetant ce qui vient d'être dit), permet de rattraper la chaîne i.

On pourrait croire qu'il y a ellipse, mais ce n'est pas vraiment le cas, puisqu'on ne saurait que faire de *ce repas* dans :

(32) Aujourd'hui, j'ai mangé [le même repas que [celui que j'ai pris hier]$_j$]$_i$. [Ce repas]$_{i?/j?/k?}$ était composé de frites.

Un autre test serait celui de la quantification. Si on peut dire qu'il y a *deux X*, alors il y aurait deux référents :

(33) * Il y a deux taux de natalité, l'un français, l'autre allemand.
(34) * Il y a deux indemnités présidentielles, l'une pour le général de Gaulle, l'autre pour F. Hollande.
(35) Il y a deux repas.

On a cependant vu que la gestion des chaînes était bien plus compliquée, et ce test, s'il peut fournir un indice, n'est pas très opératoire.

3.4.7 La visée communicative et le structuralisme

Finalement, partant de la remarque de Huang (2014, p. 224) selon laquelle il y a très peu de constance référentielle sans recours au contexte, et élaborant sur la réflexion déjà citée de Asher (1993) qui pointe l'importance de la visée communicative dans l'établissement des référents abstraits, on pourrait dire que, adaptant la distinction de Kripke entre le *semantic referent* et le *speaker's referent* , ce qui compte pour individuer les référents abstraits, c'est leur position par rapport aux autres référents.

Dans ce principe structuraliste, les limites d'un référent sont celles des autres référents, et un référent peut varier dans la mesure où les autres varient aussi. La référence serait alors le principe d'opposition qui unit le référent aux autres référents du système.

Ainsi le suicide de Werther coïncide avec la mort du Werther s'il n'est pas dans l'intention du locuteur de les séparer. Par contre, dans un article qui discuterait de la volonté de Werther de se donner la mort (volonté qui s'exprime à minuit, plusieurs heures avant la mort effective), alors il pourrait être utile de séparer les deux.

Tout comme il est utile d'opposer le taux de natalité français et le taux allemand si cela constitue le propos de l'article (par exemple, « Évolution comparée des taux de natalité français et

allemand »), mais pas si les deux valeurs sont simplement données dans une liste, sans effet d'opposition.

C'est un cas similaire que nous avons rencontré dans un article (texte T0) qui adaptait une échelle de mesure du burnout : les chercheurs adaptaient une échelle étrangère (l'échelle de Ferris) pour des sujets français. Il était alors tout naturel d'opposer les deux échelles. Mais, dans une étude qui se sert de l'échelle française et ne fait que mentionner l'échelle de Ferris pour signaler son origine, il n'y a pas d'intérêt à créer un référent, et donc une chaîne de référence, car l'échelle de Ferris originale et sa version française ne s'opposent alors pas.

Il y a cependant deux limites à une telle approche. D'une part, elle est pragmatique et non philosophique. Tout dépend donc du sens de « référence » dans « chaîne de référence ». D'autre part, elle est ou trop précise ou trop générale : elle permet sans doute, dans un petit extrait ou un paragraphe, d'opposer quelques termes entre eux et donc différentes chaînes. Elle permet aussi d'établir les grands thèmes d'un texte. Mais elle est difficilement opératoire pour suivre précisément (puisque l'annotation se fait au niveau des termes) un référent sur toute la longueur d'un texte de plusieurs milliers de mots, puisqu'il faut s'attendre à ce que les oppositions changent en fonction des parties ou des paragraphes.

3.5 Typologie des noms abstraits et des prédicats

Nous avons jusqu'à présent confondu les noms prédicatifs et les noms abstraits. Il convient maintenant d'évoquer les différents types de noms abstraits.

Certains sont prédicatifs, mais Huyghe (2014) montre qu'il existe deux définitions de « noms prédicatifs » :

1. « le N a la capacité de former, avec un V support à deux places, un constituant prédicatif »,
2. « le N peut directement régir des arguments et comporte donc une structure argumentale comparable à celles des verbes ou des adjectifs. »

La première définition est celle de Gross (2012). La seconde se rapproche plus des « noms syncatégorématiques » décrits par Kleiber (1981). Pour notre part, nous considérons que les « noms prédicatifs » correspondent à la deuxième définition.

Certains noms, dit Huyghe, appartiennent à l'une ou l'autre de ces définitions, ou bien aux deux, ou bien à aucune :

- noms prédicatifs au sens des deux définitions : *réparation, échange, traversée*,
- au sens de la première définition : *jardinage, natation, patinage* (car ils n'ont pas de structure argumentale : *Pierre fait du jardinage vs * le jardinage de Pierre*),
- au sens de la deuxième définition : *explosion, disparition, écroulement* (car ils ont une structure argumentale, mais pas d'agent : *l'explosion de la bombe*, *la disparition de l'enfant*, *l'écroulement du bâtiment* vs ** commettre une explosion*, ** effectuer une disparition*, ** accomplir un écroulement*),
- au sens d'aucune des deux définitions : *séisme, incident, festivité* (pourtant Gross (2012) considère qu'ils sont prédicatifs).

Huyghe (2012) montre qu'il existe, de même, plusieurs types de noms d'événements, qui ressemblent plus ou moins aux noms d'objets, ou plus ou moins aux noms prédicatifs.

On peut aussi montrer qu'il y a plusieurs types de noms d'actions, ou plusieurs types de noms abstraits (par exemple *la justice*, *la laïcité*) qui ne rentrent dans aucune des catégories étudiées jusque-là.

Cela nous apprend que tous les noms abstraits ne peuvent pas être traités de la même manière, et que les référents de ces noms ne sauraient être ni individués ni identifiés de la même manière. Par exemple, *jardinage* a-t-il le même référent dans les deux phrases :

(36) Pierre fait du jardinage.
(37) Paul fait du jardinage.

Car, contrairement à ce que pense Huyghe, nous croyons qu'on peut opposer *le jardinage de Pierre* et *le jardinage de Paul*, comme dans :

(38) Ah ! Le jardinage de Pierre, ce n'est pas celui de Paul !

La situation est donc bien plus compliquée que ce que cette partie a laissé entrevoir avec les seuls noms prédicatifs.

4 Conclusion

Au terme de ce parcours, on ne peut que constater que la référence des mots abstraits n'est pas acquise. Et même en faisant appel à la notion de « réification » pour décrire ces référents abstraits, il n'y a pas de moyen sûr de les suivre sur tout un texte.

La plupart des idées que nous avons exposées ne sont opératoires que sur un court extrait, ou bien sur le sens général du texte, mais ne permettent pas l'annotation fiable d'un texte *entier*. En effet, l'annotation requiert un examen détaillé des formes linguistiques, qui font apparaître les multiples façons dont le locuteur envisage un « référent abstrait ». Mais ces différentes vues se suivent et se superposent sans former une chaîne unique.

On pourrait résoudre ce problème en considérant un « référent abstrait plus ou moins flou », mais on se heurterait alors à la précision des expressions linguistiques, ce qui obligerait à faire quelques contre-sens, comme annoter un taux de natalité qui serait à la fois français et allemand, ou un repas qui serait à la fois d'aujourd'hui et d'hier.

Le problème semble venir de l'absence d'une théorie de la référence qui prennent en compte les référents abstraits. En effet, contrairement aux référents concrets, et éventuellement aux événements (que Davidson est prêt à placer à côtés des objets concrets), les référents abstraits n'ont rien d'Idées platoniciennes qui resteraient inchangées tout au long du texte. Bien au contraire, elles sont en constant mouvement, et le locuteur n'y réfèrent qu'à travers tel ou tel aspect.

On pourrait alors se demander s'il existe quelque chose comme une « chaîne de référence abstraite » qui parcourerait tout un texte scientifique. Il faudrait plutôt considérer que les « référents abstraits » n'existent pas en tant que tels, qu'il ne s'agit que de réifications qui peuvent donner lieu à quelques relations anaphoriques sur un court passage, comme un paragraphe,

mais qui en aucun cas ne durent tout le long du texte. Dans ce cas, il y aurait une succession des chaînes de « référents abstraits », mais relativement courtes.

Ou alors il faudrait redéfinir le terme « référence ».

Chapitre 2

Choix d'un corpus

Avant de réfléchir à l'annotation, il nous faut constituer un corpus d'articles de recherche au format IMRaD. Nous retenons les critères suivants :

- les textes doivent être des articles scientifiques de recheche, ce que nous définirons un peu plus loin,
- ils doivent avoir une structure IMRaD, puisque ce format impose une certaine codification dans les différentes parties, qui peuvent alors être facilement comparées, comme nous l'avons indiqué en introduction,
- ils doivent être en français, puisque notre étude porte sur les chaînes de référence en français,
- ils doivent être disponibles en version électronique[8], puisque l'annotation se fait avec un programme informatique,
- ils doivent être disponibles librement, c'est-à-dire sans nous imposer des frais supplémentaires,
- ils ne doivent pas être trop techniques, afin de faciliter le calcul de la référence des expressions que nous devons annoter.

1 Qu'est-ce qu'un article de recherche ?

Swales (1990, p. 93) décrit l'article de recherche de la façon suivante :

[8] C'est-à-dire qu'ils doivent avoir été édités directement au format numérique. C'est pourquoi le portail `persee.fr` est inutilisable, car les articles sont scannés, et la version « texte » n'est qu'une reconnaissance optique de caractères, dont on sait qu'elle n'est pas toujours très fiable et qu'elle contient beaucoup de données parasites (notamment les en-têtes et les pieds de page). Les notes de bas de page sont par ailleurs difficile à distinguer du corps du texte. Ce qui signifie que pour pouvoir utiliser un corpus à partir des articles de `persee.fr` , il faut d'abord procéder à un nettoyage et une vérification de chaque texte, ce qui manque de pratique.

> *The research article or paper... is taken to be a written text (although often containing non-verbal elements), usually limited to a few thousand words, that reports on some investigation carried out by its author or authors. In addition, the [research article] will usually relate the findings within it to those of others, and may also examine issues of theory and/or methodology. It is to appear or has appeard in a research journal or, less typically, in an edited book-length collection of papers.*

De cette description, il ressort plusieurs points. D'abord, l'article de recherche s'oppose à d'autres types de communications scientifiques (écrites ou orales). La simple lecture de la table des matières de Swales (1990) en révèle quelques-uns : « *abstracts, research presentations, grand proposals, theses and disserations, reprint requests* » (c'est-à-dire une courte lettre demandant une copie d'un article à l'auteur ou à un bibliothécaire ; ce qui se fait de nos jours par un e-mail ou un formulaire du « prêt entre bibliothèques »). Swales (2004, ch. 7) retient aussi des communications plus courtes, comme les « *conference preprints, and "notes" of various kinds that report technical innovations or observations and compilations of findings* ». On pourra aussi citer les compte-rendus et les posters (Poudat, 2006, p. 51).

Autre distinction qu'évoque Swales (1990) et qui est détaillée quinze ans plus tard (Swales, 2004, pp. 207–208) : celle, parmi les articles de recherches, entre :

- les articles expérimentaux,
- les articles théoriques,
- les articles dont le but est de proposer un état de l'art sur une question (ce qu'il appelle « *review articles* »).

Les articles théoriques sont des articles plus argumentatifs, qui ne relatent pas une expérimentation, et dont le sujet ne se prête pas au format IMRaD. Dans certaines disciplines, l'expérimentation n'est pas possible (l'astrophysique, par exemple), alors que d'autres la refusent (comme la littérature, qui pourtant pourrait trouver avantage à des études textométriques ou d'analyses statistiques textuelles).

On peut également opposer les articles de recherche, qui visent un public de chercheurs (généralement de la même discipline que l'auteur), aux articles de vulgarisation, qui visent le grand-public. La vulgarisation s'apparente « au discours journalistique » et s'oppose à « la pratique de communication scientifique académique, qui présente la particularité de faire coïncider le public de ses producteurs et celui de ses consommateurs » (Poudat, 2006, p. 51). Nous pouvons prendre l'exemple de la *Revue électronique de Psychologie Sociale*, qui, si elle est une revue à comité de lecture, publiée par une association (l'Association française de Psychologie Sociale) dont le siège est dans un département universitaire (Université Paris Ouest Nanterre La Défense), se déclare pourtant « de vulgarisation scientifique ». Les « consignes aux auteurs » du premier numéro spécifient clairement que « le style attendu est plus proche du "journalisme scientifique" que de l'habituel article empirique ». Il est en outre explicitement recommandé d'« évite[r] surtout la structure classique de la partie méthodologie d'un article empirique (participants, procédure, etc.) ».

Nous nous attacherons donc dans ce travail uniquement aux articles de recherche expérimentaux. Ceux-ci ont certaines caractéristiques spécifiques (Poudat, 2006, pp. 51 *sqq.*). En premier lieu, ils subissent des contraintes éditoriales « de forme et de contenu » :

> L'article scientifique obéit à des règles et à des codes particuliers au niveau du contenu (lourd appareillage théorique et méthodologique, présentation de résultats nouveaux ou de syn-

> thèses critiques de l'état des connaissances dans un domaine particulier du savoir) et de sa forme (importance du paratexte—notes, références bibliographiques, annexes, tableaux, schémas—, recours à un style impersonnel et utilisation d'un vocabulaire spécialisé). À travers son discours, le chercheur montre qu'il a intégré non seulement les connaissances de son domaine, mais aussi les savoir-faire, les codes, les valeurs, quand ce ne sont pas les tics (Boure, 1998, p. 107, cité par Poudat, 2006).

Ces articles répondent à une structure spécifique (Poudat, 2006, p. 53) :

- un titre,
- des noms d'auteurs et leurs affiliations,
- un résumé, parfois accompagné d'une traduction anglaise (parfois il n'y a *que* la traduction),
- une bibliographie,
- un corps, divisé en :
  - une introduction,
  - un développement (qui respecte ou non le format IMRaD),
  - une conclusion,
  - des notes de bas de pages ou de fin d'article.

Enfin, il faut noter que

> la conformité scientifique et rédactionnelle des textes est généralement validée par un comité scientifique de lecture extérieur au comité de rédaction, bien que les modalités de sélection varient selon les revues (nombre de lecteurs, anonymat des textes, etc.) et selon leurs conceptions de la pratique scientifique (Poudat, 2006, p. 52).

2 Le format IMRaD

Nous voudrions dans cette section brosser—à gros traits—le portrait du format IMRaD.

2.1 Description

Les différentes parties IMRaD ont chacune une fonction propre (Müller-Gjesdal, 2013). Deux types de références décrivent ce format :

- les articles et ouvrages scientifiques, généralement linguistiques,
- les manuels (comme le *Publication Manual of the American Psychological Association*) ou les consignes auteurs (comme le *Guide de rédaction scientifique* de la revue *Vertigo*[9]).

Si les ouvrages de référence sur le format IMRaD restent ceux de Swales (1990, 2004), nous nous appuyons, pour le reste de ce (très) rapide aperçu, sur la présentation de Müller-Gjesdal (2013).

[9] http://vertigo.revues.org/5402 , consulté le 20 janvier 2016.

Le rôle de l'introduction est de présenter l'objet d'étude, en trois « mouvements » (« *a "move" in genre analysis is a discoursal or rhetorical unit that performs a coherent communicative function* » note Swales (2004, p. 228)) :

- établir un sujet
- établir une « niche » (c'est-à-dire justifier l'étude),
- occuper la niche (c'est-à-dire décrire l'étude) (Swales, 1990).

La partie « méthodologie » (ou « méthodes ») présente le « corpus », c'est-à-dire les données utilisées pour l'expérimentation et la façon dont elles ont été recueillies, mais aussi le traitement de ces données et les éventuels instruments ou logiciels utilisés.

La partie « résultats » décrit les résultats de l'expérimentation, en notant non seulement les observations cohérentes, mais aussi les anomalies.

La partie « discussion » explique les résultats, et les confronte à ceux des études qui ont déjà été menées sur le même sujet ou sur des sujets voisins.

L'article se clôt généralement par une conclusion, dont l'existence (malgré son absence de l'acronyme « IMRaD ») est reconnue par Swales (2004), mais qui n'est pas explorée plus avant. Nous avons choisi de l'inclure comme une partie à part : notre format est donc plutôt de type « IMRaDC ».

2.2 Histoire

Pontille (2007)[10] explique que le format IMRaD s'est développé à partir du XVIIe siècle dans le but de garantir la reproductibilité d'une expérience : la structure trouve donc avant tout son intérêt dans les sciences expérimentales, notamment en physique et en chimie.

À l'époque moderne, en 1979, il a été codifié par l'American National Standards Institute (*American National Standard for the preparation of scientific papers for written or oral presentation*). C'est donc dans les publications de langue anglaise et dans les sciences de la nature que ce format se développe.

Bazerman (1988, ch. 9) décrit ensuite comment ce format s'est répandu dans les sciences humaines. La première discipline à l'adopter est la psychologie expérimentale, au moment où elle cherchait à fonder un discours propre, différent de celui de la tradition philosophique (cette tradition est encore vivace en France, c'est pourquoi, comme nous le verrons ci-dessous, le format IMRaD est rare dans la psychologie française). C'est donc la psychologie expérimentale qui a fixé les standards, et c'est pourquoi le manuel de style de référence reste le *Publication Manual of the American Psychological Association* , dont Bazerman décrit l'évolution depuis sa première publication en 1929 (pp. 261 *sqq.*).

[10] On trouvera une histoire plus détaillée dans Swales, 1990, pp. 110–117.

3 Présence dans la recherche française. Sélection des revues

3.1 Le paysage français

Puisque le monde anglo-saxon semble être un modèle pour la production scientifique du monde entier, on s'attendrait à ce que le format IMRaD ait pénétré la publication scientifique française. C'est vrai en ce qui concerne les sciences de la nature... mais le modèle anglo-saxon a été tellement prédominant que la production française dans ces disciplines est écrite désormais directement en anglais. Jacques (2013) souligne que certains articles, du traitement automatique des langues (TAL) notamment, se rapprochent de ce format, mais les exemples qu'elle donne montrent que s'ils respectent l'idée d'une progression d'une exposition en étapes progressives, ils ne se conforment pas strictement au format IMRaD.

Or les sciences humaines n'utilisent que très peu ce format. L'exemple de la géographie est un exemple emblématique de cette répartition des publications dans la recherche française. La géographie humaine (démographie, aménagement urbain, flux migratoires, etc.) est une science humaine dont le portail `cairn.info` , portail de revues en SHS[11], offrent de nombreuses revues. Les articles y sont en français, mais ne respectent pas le format IMRaD. La géographie physique (géomorphologie, hydrologie, pédologie, etc.), par ailleurs, est une science de la terre, donc une science de la nature, dont les revues sont indexées, non dans les portails de SHS, mais dans les bases (anglaises) de sciences « dures » comme le *Web of Science* ou *ScienceDirect* . Les articles ont souvent la structure IMRaD, mais sont en anglais, même quand ils émanent de chercheurs français travaillant dans des centres de recherche en France. Il existe cependant des exceptions, comme la revue *Géomorphologie*, une revue de géographie physique qui publie quelques articles IMRaD en français. Mais la plupart des articles publiés, même dans cette revue pourtant française, sont en anglais.

La sociologie, elle, offre un exemple de la répartition des articles de structure IMRaD à l'intérieur d'une science humaine (Pontille, 2003) : si un certain nombre d'articles utilisent ce format, notamment dans les analyses quantitatives, la quasi-totalité de ces articles IMRaD sont en anglais...

Quant aux humanités[12], ils n'utilisent quasiment jamais le format IMRaD, comme le confirme Rinck (2006, p. 158), qui étudie des articles de lettres et de linguistique : « un examen rapide des articles de notre corpus suffit à s'en convaincre : une telle structure [= IMRaD] apparaît très rarement. On l'observe sous une forme plus ou moins canonique dans 7 articles de SCL [= SCiences du Langage] (soit une proportion de 6,4 % des articles du corpus), et jamais en LET [= Lettres] ». En effet, en lettres

> prédomine un profil de textualité qui serait en quelque sorte le contre-pied du modèle IMRD et plus largement de la subdivision du texte et du guidage du lecteur en vigueur en sciences ; la recherche sur les textes littéraires se présente comme une approche unifiée, peu propre à être caractérisée au moyen des divisions pré-établies de type « méthode », « résultats », « analyse », etc. (p. 163).

[11] Sciences humaines et sociales.
[12] Art, lettres et langues.

3.2 Recherche en bibliothèque

Notre premier réflexe, pour trouver des articles, a été de nous tourner vers le corpus Scientext (Tutin et Grossmann, 2013[13]), mais, d'après ce qu'on peut voir sur le site Internet, les quelques articles de psychologie ne sont pas au format IMRaD.

Pour tenter, néanmoins, de trouver un corpus d'articles de recherche *en français* et *respectant le format IMRaD*, nous avons décidé d'examiner un maximum de revues en français, toutes disciplines confondues. Nous avons simplement passé en revue les articles du dernier numéro (parfois des quelques derniers numéros) de *toutes* les revues sur papier dans les bibliothèques de psychologie (ce qui inclut les sciences de l'éducation) et de sociologie (ce qui inclut l'ethnologie), les deux disciplines où nous étions le plus susceptible de trouver des articles au format IMRaD. Puis nous nous sommes tourné vers le portail `cairn.info`, puisqu'il ne semble pas exister de portail équivalent *en français* pour les sciences de la nature. Nous avons ainsi parcouru tous les articles des deux derniers numéros de trois à cinq revues de chacune des disciplines proposées (sauf les arts, le droit, l'histoire, la philosophie, ainsi que le sport, qui ne contient que trois revues) sur la page d'accueil du site.

Plusieurs points peuvent être dégagés de cette recherche. D'abord, comme Pontille (2003) l'annonçait, les sociologues français n'usent qu'exceptionnellement du format IMRaD. Ensuite, comme nous l'a expliqué une bibliothécaire de la bibliothèque de psychologie, la psychologie clinique française est d'orientation lacanienne et jungienne, ce qui signifie qu'elle n'est pas dans une démarche expérimentale (mais plutôt philosophique), et ne produit donc pas d'articles expérimentaux susceptibles de respecter le format IMRaD. Cependant, la psychologie sociale et les sciences de l'éducation semblent avoir une majorité d'articles IMRaD, dont quelques-uns sont *en français*. Par ailleurs, la revue *Enfance*, qui s'intéresse aux sciences du développement, contient également beaucoup d'articles IMRaD en français, mais elle est isolée en ce sens que toutes les autres revues de cette discipline sont en anglais.

Il faut souligner qu'on trouve des articles de format IMRaD dans toutes les autres disciplines que nous avons examinées, mais il sont très rares. En psychologie sociale, par contre, ils sont la norme.

Cependant, les revues repérées sur papier en bibliothèque ne nous étaient pas forcément accessibles en ligne (car ne comptant pas parmi les abonnements de l'Université de Strasbourg ou de la Bibliothèque Nationale et Universitaire de Strasbourg). Nous avons tout de même récupéré un article de la *Revue Internationale de Psychologie Sociale* : ce sera le texte T0 de notre corpus.

Ensuite, nous avons opté pour une approche plus systématique, en examinant systématiquement les articles de toutes les revues en libre accès sur le portail `revues.org`.

[13] `http://scientext.msh-alpes.fr`

3.3 Étude systématique d'un corpus

Le portail revues.org (OpenEdition) propose nombre d'articles récents et directement en version HTML. Il y a parfois un « mur mobile », c'est-à-dire un délai d'attente pour accéder à la version « en texte intégral », mais ce délai n'est généralement que de deux ans.

Nous avons étudié systématiquement les articles des quatre derniers numéros disponibles « en texte intégral » (c'est-à-dire sans avoir besoin de souscrire à un quelconque abonnement) des 420 revues disponibles sur le portail à la date du 15 janvier 2016.

Un certain nombre de revues ont dû être retirées du corpus :

- celles en langue étrangère (anglais et espagnol, notamment),
- celles dont la liste des numéros n'est pas normalisée et ne peut donc pas être facilement récupérable de façon automatisée. Cela ne pose problème que pour la revue *Vertigo*, citée par Jacques (2013) comme exemple de revue dont les articles ont souvent la structure IMRaD. Elle n'apparaît donc pas dans le corpus suivant, mais il faudra tout de même la prendre en compte dans les résultats,
- celles dont le ou les liens d'accès au numéro sont brisés ou ne répondent pas (le problème vient du site d'OpenEdition),
- celles qui ne sont pas des revues. En trouve en effet sur revues.org un certain nombre de ressources qui n'ont pas, paradoxalement, le format d'une revue. Par exemple, le corpus Eve, pour « Émergence des Vernaculaires en Europe »[14], est un corpus de textes anciens pour l'étude des langues vernaculaires, et n'est pas du tout une revue.

Il reste donc 341 revues. Ce sont des revues SHS qu'OpenEdition considère comme scientifiques. Le dépouillement des quatre derniers numéros pour chacune d'elles (ou moins si la revue est très récente et n'a pas encore publié quatre numéros) donne un corpus de 21 689 articles. Certains de ces articles n'ont pas d'intertitres, soit parce que ce sont des formats qui n'en usent traditionnellement pas, comme les comptes-rendus d'ouvrage, soit parce que l'auteur n'a pas choisi d'en faire usage (ce qui est notamment le cas en littérature, où près d'un article sur deux n'a pas d'intertitres). Nous avons exclu ces articles. Le corpus comprend donc 12 367 articles avec intertitres.

Parmi ces articles, nous avons cherché ceux qui contenaient, dans leurs intertitres, les termes « méthode », « méthodologie », « méthodologique », « expérimentation » ou « protocole », ainsi que « résultat » et « discussion », au singulier et au pluriel. Comme nous ne pouvions pas déterminer de façon automatique la langue de l'article, et afin de ne pas pénaliser les revues françaises qui publient des articles en anglais, nous avons aussi cherché les termes en anglais. En fait, seules trois revues, *ArcheoSciences*, *Géomorphologie* et *Recherche et pratiques pédagogiques en langues de spécialité* ont un nombre d'articles IMRaD en anglais significativement élevé (entre 30 et 56 %). Mais ce sont aussi les revues qui contiennent le plus d'articles IMRaD en français, par comparaison avec les autres revues du corpus.

Seuls 329 plans d'articles contiennent les mots-clés cherchés. Afin d'éviter les « faux-positifs », c'est-à-dire les articles qui ont les mots-clés cherchés dans leurs inter-titres, mais n'ont pas de structure IMRaD, nous avons examiné manuellement tous ces résultats. Nous en avons ainsi éliminé six. La discussion, par exemple, peut être le « phénomène étudié » (« discussion en pe-

[14] http://eve.revues.org

tits groupes »). La méthode peut être ce qui est discuté (« choix des méthodes de cartographie et des modes de gestion de l'aléa d'inondation »). Les résultats, enfin, peuvent n'être que des « exemples de résultats ».

Il reste donc 323 « vrais » articles IMRaD, dont 255 en français, soit 2.61 % (2.06 % en français), dans 78 revues (les autres n'ont aucun article IMRaD dans leur quatre derniers numéros).

En se servant des catégories disciplinaires établies par OpenEdition, on note d'abord que seules trois disciplines ont un « taux IMRaD » supérieur à 10 % : la psychologie (18.6 %), l'éducation (12.7 %) et le management (11.1 %).

Les disciplines littéraires ne comptabilisent aucun article IMRaD, et les « arts et humanités » ont un taux extrêmement faible : 0.37 %.

La comparaison par disciplines rencontre cependant des limites importantes. D'abord, la plupart des revues sont catégorisées dans plusieurs disciplines à la fois. Ensuite, il y a des écarts importants dans le nombre de revues par discipline (par exemple 134 revues en « arts et humanités », mais seulement deux en « sciences de la santé et de la santé publique »). Enfin et surtout, le choix de certaines catégorisations est douteux, comme *Géomorphologie* (qui est une revue de géographie physique) en « histoire et archéologie » et en « management et administration », ou la *Revue de Primatologie* en... « éducation ».

Nous avons donc préféré nous intéresser aux revues individuellement. Sur les 78 revues qui ont des articles IMRaD dans les quatre derniers numéros, nous avons choisi d'étudier plus avant celles qui en ont plus de 15 %, soit 20 revues. Pour chacune d'elles, nous avons dépouillé l'ensemble des articles avec intertitres des dix derniers numéros « en texte intégral » (ou moins si la revue est trop récente pour avoir déjà dix numéros), soit 1580 articles. Sur ces articles, 387 ont une structure IMRaD (314 en français), soit 24.49 % (19.87 %).

Que ce soit en étudiant les quatre ou les dix derniers numéros, l'ordre de ces revues en fonction de la fréquence d'articles IMRaD est globalement le même :

1. *Géomorphologie*, avec près de 58 % d'articles IMRaD, est largement en tête du classement. Néanmoins, sur 40 articles IMRaD, seuls 25 sont en français.
2. La plupart des autres revues ont entre 11 % et 44 % d'articles IMRaD. Les disciplines tendent à se répartir entre les sciences dures (comme *Physio-Géo* et *ArcheoSciences*, qui est une revue qui cherche à résoudre des problèmes archéologiques à l'aide des sciences dures) et les sciences de l'éducation, auxquelles on doit rajouter diverses disciplines (la linguistique, la sociologie ou le management) qui n'ont chacune qu'un ou deux représentant.
3. En dernière position se trouve *Méditerranée*, seule revue à avoir moins de 10 % d'articles IMRaD.

Quinze de ces vingt revues ont un taux plus élevé d'articles IMRaD dans les quatre derniers numéros que dans les dix derniers. C'est par exemple le cas pour la dernière revue mentionnée, *Méditerranée*, qui passe de 9.7 % à 19.15 %. Cela montre que le format IMRaD pénètre de plus en plus les publications en français.

Sur le plan de la langue, seules trois revues ont un nombre significativement élevé d'articles IMRaD anglais :

- *ArcheoSciences* (41 en anglais contre 32 en français, ce qui signifie que 56 % des articles IMRaD sont en anglais),
- *Géomorphologie* (15 contre 25, soit 37.5 %),
- *Recherche et pratiques pédagogiques en langues de spécialité* (5 contre 12, soit 29.4 %).

Les autres revues ont moins de trois articles IMRaD (et souvent aucun) en anglais.

Néanmoins le nombre d'articles IMRaD pour ces trois revues reste conséquent, et même supérieur, pour *ArcheoSciences*, à celui des autres revues qui publient majoritairement en français. On peut se demander s'il n'y a pas là un effet de « contamination » de cette structure d'origine anglo-saxonne : le format IMRaD se diffuserait plus rapidement dans les articles en français si ceux-ci sont publiés dans une revue bilingue. Néanmoins, il y a trop peu de données pour tirer une conclusion, d'autant plus que nous ne pouvons pas calculer les fréquences relatives pour l'anglais et le français, ne pouvant déduire la langue de *tous* les articles (mais seulement celle des articles IMRaD).

Du point de vue de la fréquence absolue d'articles IMRaD en français, la plupart des revues permettent d'espérer au moins un article de ce type par numéro. Six revues permettent d'en espérer deux ou plus. C'est *ArcheoSciences* qui offre le plus d'articles IMRaD en français (32 sur les dix derniers numéros, soit trois articles par numéro en moyenne).

En ce qui concerne le caractère scientifique de ces revues, toutes se veulent scientifiques, s'adressent à des chercheurs (à lire la présentation de chacune d'elle), ont un comité scientifique (visible sur leur site internet) et la plupart ont une « phase d'évaluation en double lecture aveugle »[15]. Nous pensons donc que toutes répondent aux critères de scientificité.

Pour finir, nous proposons de classer ces revues en trois groupes disciplinaires :

- les sciences dures (nous y incluons l'archéologie, puisque les articles IMRaD qu'on y trouve sont surtout des articles concernant des méthodes issues des sciences dures), qui fournissent 45 % des articles IMRaD (35 % en français) :
  - *ArcheoSciences*,
  - *Physio-Géo*,
  - *Géomorphologie : relief, processus, environnement*,
  - *Revue de primatologie*,
  - *Méditerranée*,
  - *Préhistoires méditerranéennes* ;
- les sciences de l'éducation, qui fournissent 38 % des articles IMRaD (45 % en français) :
  - *Travail et formation en éducation* (arrêtée depuis 2011),
  - *L'Orientation scolaire et professionnelle*,
  - *Questions Vives*,
  - *Recherches et éducations*,
  - *Revue internationale de pédagogie de l'enseignement supérieur*,
  - *RDST*,
  - *Recherche et pratiques pédagogiques en langues de spécialité*,
  - *Revue française de pédagogie*,
  - *Distances et médiations des savoirs*,

[15] Politique éditoriale de *Questions Vives* http://questionsvives.revues.org/830 , consulté le 16 janvier 2016

  - *Les dossiers des sciences de l'éducation* ;
- les autres disciplines (linguistique, management, sociologie), qui fournissent 17 % des articles IMRaD (20 % en français) :
  - *Finance Contrôle Stratégie*,
  - *Acquisition et interaction en langue étrangère*,
  - *TIPA. Travaux interdisciplinaires sur la parole et le langage*,
  - *Perspectives interdisciplinaires sur le travail et la santé*.

Nous proposons donc de prendre comme corpus pour l'étude des chaînes de référence des articles des sciences dures et des sciences de l'éducation, à proportion égale, parmi les revues citées ci-dessus (en rajoutant aussi la revue Vertigo, que nous avons dû exclure de cette recherche parce que son format ne permettait pas de récupérer la liste des numéros automatiquement, mais qui, après analyse manuelle, s'avère contenir plusieurs articles IMRaD), notamment *ArcheoSciences*, qui présentent le double avantage d'avoir le plus grand nombre d'articles IMRaD en français, mais aussi de « trait[er] de l'application de diverses techniques scientifiques (sciences physiques, chimiques, mathématiques, sciences de la terre et de l'univers, et sciences de la nature et de la vie) à la résolution de problématiques archéologiques »[16]. Les articles sont donc, pensons-nous, à destination des archéologues, qui ne sont pas des spécialistes des sciences dures, ce qui laisse supposer un moindre degré de technicité dans les démonstrations et le vocabulaire.

À terme, les autres disciplines pourraient fournir une sorte de corpus de contrôle : nous pourrions alors vérifier si nos analyses peuvent être étendues aux articles des autres sciences humaines, comme la linguistique, la sociologie ou le management.

4 Sélection des textes

4.1 Problèmes...

Maintenant que nous avons des revues, accessibles en ligne, où chercher des articles de format IMRaD en français, nous devons encore choisir quels articles sélectionner. Car là aussi, il y a des contraintes, dont la principale concerne les intertitres.

Comme le dit Pontille (2003, p. 58), le format IMRaD « constitue... une structure générique qui se retrouve assez peu dans les textes sous sa forme pure ». C'est pourquoi il convient d'examiner la structure des articles des revues retenues avant de les inclure dans le corpus.

Jacques (2013, p. 204) identifie deux types d'intertitres :
- l'intertitre « métatextuel », « qui explicite la nature de la section qu'il couvre, sans interaction avec le texte de cette section ». Par exemple, les intertitres *Introduction, Résumé, Bibliographie*,
- l'intertitre « textuel », qui « donne à voir la thématique qui va être développé dans la section et participe à la construction du propos du texte ».

[16] http://archeosciences.revues.org/

Dans d'autres recherches (par exemple Jacques, 2005), elle remarque que les titres peuvent avoir une « implication référentielle » et avoir une incidence sur les chaînes de référence.

De son côté, Pontille (2007, pp. 7–8) note que « dans le cadre de la standardisation en sections, la cohérence de la démarche prend directement corps dans la matérialité des titres du texte ». Ainsi, « les titres standardisés du texte séparent nettement chaque niveau du récit expérimental tout en assurant une unité à l'argumentation ». L'auteur peut alors « se passer de toute transition argumentée entre les différentes sections de l'article », notamment des phrases qui « apportent leur renfort de cohérence », ce qui reprend ce que disait déjà Bazerman (1988, p. 260) : « *The prescribed form of fixed sections with fixed titles creates disjunctions between mandatory sections: the author does not have to establish overt transitions and continuity among the parts* ».

Puisque les titres « métatextuels » (selon la terminologie de Jacques, 2013) non seulement sont dépourvus d'un risque d'implication référentielle qui aurait une incidence sur les chaînes de référence, mais en plus évitent l'encombrement du texte avec des transitions, qui elles aussi, peuvent avoir une incidence malencontreuse sur l'étude des chaînes, et ce d'autant que les transitions peuvent se trouver en fin ou en début de partie, selon le choix stylistique de l'auteur, nous proposons d'écarter les textes dont les intertitres ne sont pas « métatextuels ».

Or très peu d'articles respectent le format IMRaD à la lettre. Les premiers sondages que nous avons effectués ont montré que les introductions sont parfois très développées, avec des intertitres intermédiaires (de type « textuels »), qu'il peut y avoir plusieurs expérimentations et donc plusieurs parties « méthodologie » et « résultats », qu'il y a souvent une conclusion, etc.

Nous avons donc décidé de choisir des textes dont la structure IMRaD est la plus claire possible, c'est-à-dire qui présentent tous des intertitres contenant les termes « méthodologie », « résultats », « discussion », ainsi que « conclusion », et éventuellement « introduction » (qui n'est pas indispensable). Les autres inter-titres éventuels sont ignorés. Si un article a plusieurs expérimentations, et que chacune a une partie « méthodologie », alors le texte aura plusieurs parties « méthodologie ».

Nous avons également choisi les textes en fonction de leur faible technicité, c'est-à-dire que nous avons évité les textes trop techniques.

4.2 Liste des textes

Avec ces contraintes, nous avons retenu, pour les études exploratoires que nous présenterons ci-dessous (chapitres 5 et 6), cinq textes. On trouvera les références précises et le résumé de ces textes dans l'annexe A. Il suffit ici de dire que trois des textes sont orientés vers les sciences humaines :

- le texte T0, « Le rôle modérateur des compétences politiques sur le burnout », appartient à la psychologie sociale. Il n'a été annoté que pour la première étude,
- le texte T2, « La propension à l'interdisciplinarité des étudiants en situation d'innovation » et le texte T4, « Représentations socio-professionnelles et choix de la spécialisation : le cas de la filière vétérinaire rurale », appartiennent plutôt aux sciences de l'éducation.

Les deux autres techniques sont issus de revues d'archéologie, leur méthodologie (analyses physiques et chimiques) les rapproche plutôt des sciences de la nature. Ce sont les textes :

- T1 : « Les pratiques de subsistance de la population Néolithique final de la grotte I des Treilles (commune de Saint-Jean-et-Saint-Paul, Aveyron) »,
- T3 : « Caractérisation de phénomènes anthropiques par la mesure de paramètres magnétiques sur surface décapée : Premiers résultats sur le projet Canal Seine-Nord Europe ».

Chapitre 3

Construction d'un outil d'annotation

1 Introduction

Nous abordons maintenant une partie qui sera consacrée à l'ingénierie de l'annotation et de l'analyse des données. Mais il nous semble bon, dans un premier temps, de donner un premier aperçu de l'ensemble du processus d'annotation et d'analyses des données. Nous décrirons ensuite les besoins que nous avons eus pour mener à bien cette annotation et ces analyses. Nous verrons comment deux logiciels d'annotation, Glozz et Analec, y répondent, mais aussi pourquoi ils n'ont pas été suffisants et pourquoi nous avons dû programmer de nouveaux scripts. Dans la partie suivante, nous aborderons les questions méthodologiques sous un autre angle : celui de la linguistique. Nous exposerons et justifierons alors le schéma d'annotation retenu, et nos choix concernant la délimitation des maillons. Enfin, nous présenterons les problèmes spécifiques à l'annotation des référents abstraits au cours de l'exposé de notre première étude exploratoire.

Nous essayerons de rester ici au niveau méthodologique, et de raisonner en termes de besoin pour l'annotation et l'analyse. En ce qui concerne le niveau technique de l'implémentation, nous avons inclus, en annexe de ce travail, un petit « guide d'utilisation » de l'interface graphique d'annotation que nous avons écrite. Nous n'avons pas rédigé un tel guide pour les scripts d'analyse quantitative, mais nous tenons à la disposition du jury l'ensemble des codes.

2 Vue d'ensemble du processus d'annotation et d'analyse

Étape 1. Sélection du texte dans le corpus de textes IMRaD (voir chapitre 2).

Étape 2. Préparation du texte afin d'avoir un texte brut[17] prêt pour l'annotation. Cette étape comporte plusieurs opérations :

1. Copie du texte depuis le site Internet de l'éditeur.
2. Suppression des éléments hors-textes (notamment les notes de bas de page, illustrations, tableaux, liens pour le téléchargement des graphiques et tableaux, bibliographie, informations supplémentaires sur les auteurs, les droits d'auteur, etc.).
3. Ajout des métadonnées générales (par exemple le titre, la source, les auteurs, la revue, etc.) et d'identification pour nos scripts (un identifiant propre à chaque texte)
4. Ajout des métadonnées spécialisées, notamment les séparateurs de parties, l'identification des types de parties IMRaD (introduction, méthodologie, résultats, discussion, conclusion), qu'elles soient générales, ou spécifiques à telle ou telle expérimentation. En effet, comme nous l'avons expliqué plus haut (page 40), certains textes ont plusieurs parties « introduction », « méthodologie », etc. ; par exemple, si le texte présente une étude qui a nécessité deux expérimentations, il y aura une introduction générale, puis une introduction et une partie « méthodologie », voire aussi une partie « résultats », pour chaque expérimentation.
5. Marquage spécifique des inter-titres. Puisque nous n'avons pas exclu les inter-titres, mais qu'il nous a paru bon de pouvoir les repérer, tant pour l'analyse des données que pour leur affichage correct dans l'interface d'annotation, nous avons marqué chacun des inter-titres comme un paragraphe spécial. Nous avons aussi noté le niveau hiérarchique afin de pouvoir prendre en compte cette donnée plus tard.
6. Identification des tokens étranges. Puisque nos scripts utilisent le token comme unité minimale, et non le caractère (comme Glozz ou Analec), nous avons dû choisir une définition stricte du token pour pouvoir les repérer automatiquement. Mais les textes scientifiques utilisent parfois des symboles spéciaux, par exemple « δ13C » (T1), « isotopes en carbone 13 », qui est un aggrégat d'une lettre grecque, de plusieurs chiffres et d'une lettre latine. Il suffit d'indiquer dans les métadonnées les tokens un peu spéciaux de ce genre pour qu'ils soient reconnus automatiquement.

Étape 3. Analyse du contenu textuel, notamment avec TreeTagger et AntConc (voir le chapitre 5) pour déterminer les référents à annoter pour la première étude exploratoire (cette étape n'a pas lieu d'être pour la deuxième étude, puisque tous les référents qui initient des chaînes ont été systématiquement annotés).

Étape 4. Annotation avec notre script (interface graphique), en deux phases :

- Délimitation des maillons. Nous avons essayé de délimiter les expressions nominales ou pronominales avec un *chunker*, c'est-à-dire un outil qui recherche les SN et construit donc les maillons automatiquement. Nous avons traité les textes de la première étude avec un tel outil[18], en filtrant les SN dont nous avions besoin (puisque nous n'avons annoté qu'une vingtaine de référents par texte pour cette étude). Nous nous sommes cependant aperçu que, si l'utilisation d'un *chunker* permet d'économiser (dans l'hypothèse où il ne fasse pas trop d'erreurs) le nombre de clics, la charge cognitive pour contrôler les délimitations

[17] C'est-à-dire sans aucun formatage comme des mots en gras, en italiques, etc.

[18] Après avoir essayé SEM, de Tellier : `www.lattice.cnrs.fr/sites/itellier/SEM.html` , TTL entraîné par Longo : `www.racai.ro/en/tools/text/` , TreeTagger : `www.cis.uni-muenchen.de/~schmid/tools/TreeTagger` et un ensemble de règles avec Unitex : `www-igm.univ-mlv.fr/~unitex/` , nous avons opté pour ce dernier.

était en fait plus importante que celle pour simplement annoter à partir du texte brut. De plus, le travail du *chunker* occupe visuellement l'espace et empêche souvent de voir ses oublis (nous avons constaté que nous faisions beaucoup plus d'erreurs avec un *chunker* que sans). Sans compter le fait qu'on a tendance à trop lui faire confiance. Aussi, nous n'avons pas utilisé de *chunker* dans la deuxième étude exploratoire.

- Choix des valeurs pour chaque propriété de chaque maillon (c'est-à-dire noter, pour chaque maillon, sa catégorie grammaticale, sa fonction, ses expansions éventuelles, etc.). Nous avons fait cela propriété par propriété, car cela nous semble plus rapide et demander une moindre charge cognitive.

Cette démarche en deux temps nous a permis d'avoir une idée des éventuels problèmes à caractère linguistique dès la phase de délimitation des maillons, et de faire des choix mieux informés lors de l'élaboration du schéma d'annotation (construit, donc, entre les deux phases). C'est également la démarche adoptée pour le projet MC4 (Mélanie-Becquet et Landragin, 2014) et pour le projet Democrat.

Étape 5. Analyse quantitative avec nos scripts Perl, qui produisaient une grande quantité de fichiers CSV[19]. Chacun de ces fichiers contenaient des données (statistiques, fréquences, etc.) pour le corpus, chaque texte, chaque partie, chaque paragraphe, chaque type de partie, et ceci pour chaque groupe de chaînes et pour chaque chaîne.

Étape 6. Analyse qualitative en consultant les annotations avec l'interface graphique, ou bien chacun des fichiers CSV. En général, nous avons adopté une approche « descendante » : lorsque nous détections un phénomène intéressant lors de l'analyse quantitative, nous cherchions à l'expliquer en regardant dans le détail des chaînes, d'abord dans les nombreux fichier CSV, ce qui nous permettait de vérifier des statistiques ou des fréquences sur des étendues plus réduites du texte (partie, paragraphe, ou type de partie) pour un groupe de référents ou un référent unique (ou même une sélection *ad hoc* de référents). Si cela se justifiait, nous regardions enfin le texte, soit dans le concordancier que nous avons programmé, soit dans l'interface graphique afin de faire une analyse de détail.

C'est à peu près la méthode que préconise Landragin (2016) :

> Notre méthodologie suit trois phases... : la première phase relève de statistiques générales sur le texte et sur la répartition des chaînes dans ce texte ; la deuxième phase relève de l'étude approfondie—visualisation, exploration et calculs statistiques—de la suite des références du texte ; la troisième phase relève de l'étude approfondie—visualisation, exploration et calculs statistiques—des chaînes de coréférences.

[19] CSV signifie *comma separated values* , ou « valeurs séparées par des virgules ». Il s'agit d'un format de fichier qui permet d'enregistrer des tableaux : chaque séparateur de colonne est représenté par une virgule, et chaque nouvelle ligne par un retour chariot. Ces fichiers, dont le format est très simple, sont lisibles par des tableurs tels que Microsoft Excel ou LibreOffice Calc, programmes qui permettent de faire ensuite des calculs statistiques sophistiqués, des diagrammes, etc.

3 Présentation de Glozz et d'Analec

De nombreux logiciels d'annotation existent (on en trouvera une liste relativement exhaustive dans Fort, 2012, et on se reportera à Landragin (2014) pour ce qui concerne les chaînes de référence), mais nous en retiendrons ici seulement deux :

- Glozz (Widlöcher et Mathet, 2009, 2012), qui a été utilisé dans le cadre de l'annotation des chaînes de référence par Landragin (2011) et dans le cadre du projet ANCOR (Muzerelle, A. Lefeuvre, Schang *et alii*, 2014),
- Analec (Victorri, 2011), dont le nom signifie « ANAlyse de L'ÉCrit » (Mélanie-Becquet et Landragin, 2014, p. 118), qui a été utilisé dans le cadre du projet MC4 (« Modélisation Contrastive et Computationnelle des Chaînes de Coréférence ») (voir notamment Landragin, 2011) et est actuellement utilisé (et développé) dans le cadre du projet Democrat.

Il faut évoquer, avant d'aller plus loin, quelques problèmes terminologiques. Les concepteurs de Glozz, ainsi que Landragin (2014), parlent de « structure de traits » (*feature-set*) pour désigner l'ensemble des propriétés à annoter. Ils appellent donc « traits » ce que nous appelons « propriétés », et chacun de ces « traits » (par exemple la catégorie grammaticale) peut prendre certaines « valeurs » (*feature-name, feature-value*) (sujet, complément, etc.), ce que nous appelons soit « valeurs », soit « classes » (en référence au calcul statistique). Mélanie-Becquet et Landragin (2014) et Landragin (2016) semblent avoir abandonné le terme de « traits » pour utiliser celui de « propriétés ».

3.1 Glozz

Glozz (voir figure 2) a une vocation généraliste (Fort, 2012, p. 230), c'est-à-dire qu'il n'est pas orienté spécifiquement vers l'annotation des chaînes de référence, mais la façon dont il représente les annotations le rend bien adapté à une telle tâche. En effet, il utilise un modèle dit « URS », pour « unité-relation-schéma » (Widlöcher et Mathet, 2012, Fort, 2012, p. 230) : les « données [sont] annotées sous la forme d'"unité" (suite consécutive de caractères, délimitée dans une donnée unitaire que le linguiste va pouvoir enrichir d'annotations), de "relation" (lien entre deux unités) et de "schéma" (ensemble structuré d'unités, de relations, et, de manière récursive, de schémas) » (Landragin, 2014).

Voici comment ces trois éléments peuvent être combinés pour annoter une chaîne de référence :

> Une mention ou expression référentielle correspond à une « unité », qui porte (éventuellement) des annotations sous la forme d'une structure de traits, c'est-à-dire sous la forme de couples attribut-valeur. Une relation anaphorique entre deux mentions peut se construire *via* une « relation », elle-même pouvant porter des annotations sous la forme d'une structure de traits. Une chaîne de coréférence peut se construire soit *de facto* par un ensemble d'unités et de relations qui forment une liste chaînée, soit *via* un « schéma » qui groupe alors mentions et relations, et auquel on peut bien entendu adjoindre sa propre structure de traits. (Landragin, 2014)

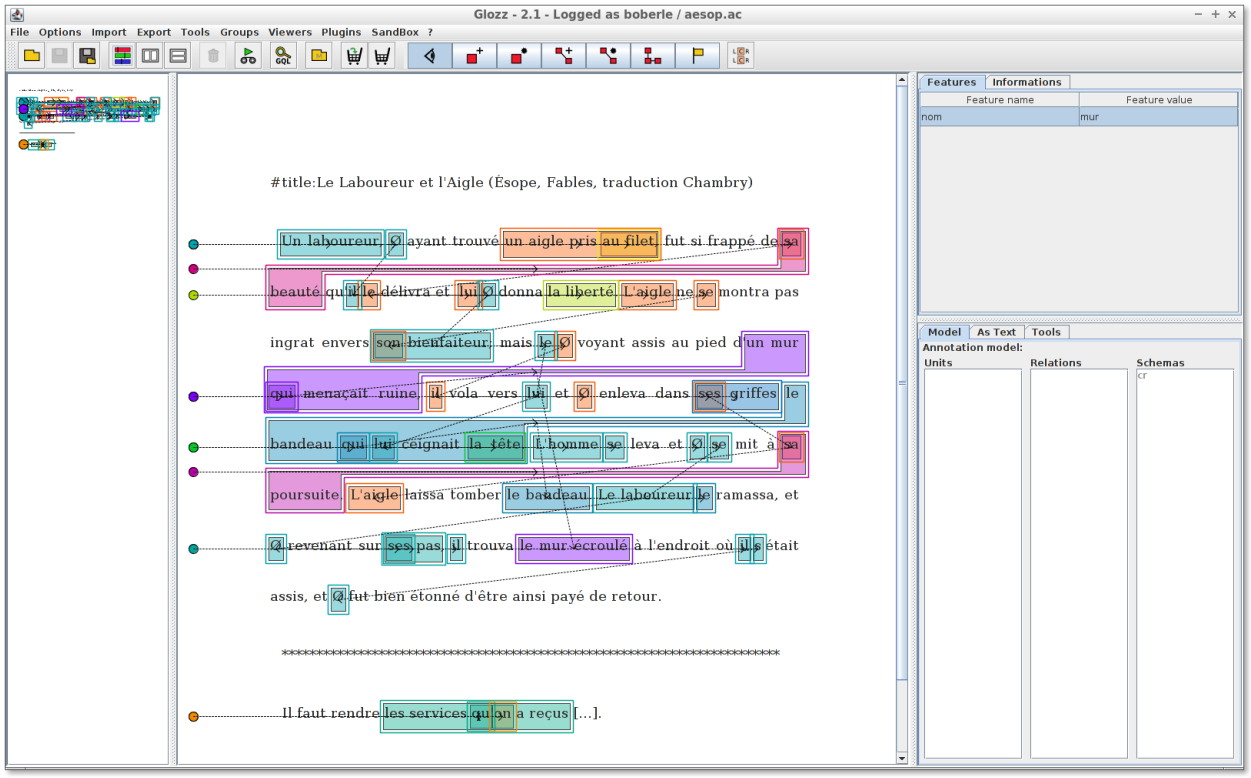

Fig. 2: Interface de Glozz.

Un langage de requête puissant, GlozzQL (pour Glozz *Query Language*), permet de faire des recherches complexes sur les annotations, bien qu'il soit « lourd à manipuler » (Fort, 2012, p. 230).

Concernant le stockage des annotations, Glozz utilise des annotations « débarquées » (Habert, 2005, p. 32) ou « déportées » (Fort, 2012, p. 230), c'est-à-dire qu'elles sont sauvées dans un fichier séparé du texte lui-même. Le corpus est stocké sous forme brute, dans un format lisible par un humain, alors que les annotations sont sauvées en XML, format lisible par un humain et par d'autres programmes (dont Analec).

« Visuellement, l'outil est très agréable et confortable, montrant toutes les imbrications et relations complexes de façon claire », notamment parce qu'il permet « de masquer certaines informations pour une tâche donnée » (Fort, 2012, p. 230). Cependant, cet affichage demande beaucoup de ressources (du moins les bibliothèques Java sur lesquelles il est construit demandent beaucoup de ressources), si bien qu'il devient rapidement inutilisable. Nous avons ainsi tenté d'ouvrir nos textes annotés avec Glozz, mais avons dû renoncer à travailler avec, devant la lenteur de chaque opération. De plus, la représentation des schémas sous forme de liens visibles *sur* le texte rend ce dernier rapidement illisible.

C'est pourquoi nous n'avons pas eu l'occasion d'utiliser Glozz dans les conditions réelles d'une annotation de chaîne de référence. Tout juste avons-nous pu tester ses fonctionnalités sur de petits textes.

3.2 Analec

D'abord conçu comme un outil qui permet à la fois l'annotation, la visualisation et l'interrogation des données (Landragin, Poibeau et Victorri, 2012), Analec (voir figures 3 et 4) est un outil qui a été « retravaill[é] pour le projet MC4 » (Mélanie-Becquet et Landragin, 2014, p. 118), c'est-à-dire pour l'annotation des chaînes de référence, pour lesquelles il dispose de fonctionnalités spécifiques (voir notamment Landragin, 2011, 2014). Il est donc plus orienté que Glozz, sans devoir le « rédui[re]... à un outil d'annotation des phénomènes de coréférence » (Mélanie-Becquet et Landragin, 2014, p. 133).

C'est le logiciel utilisé par les membres du projet Democrat, et c'est à cette occasion que nous avons eu l'occasion d'y être formé et de travailler avec lui.

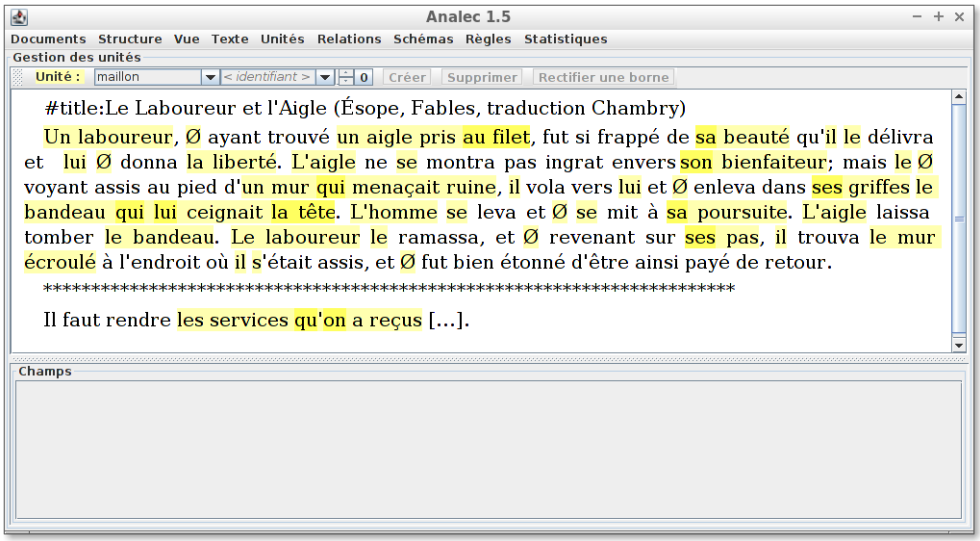

Fig. 3: Interface d'Analec.

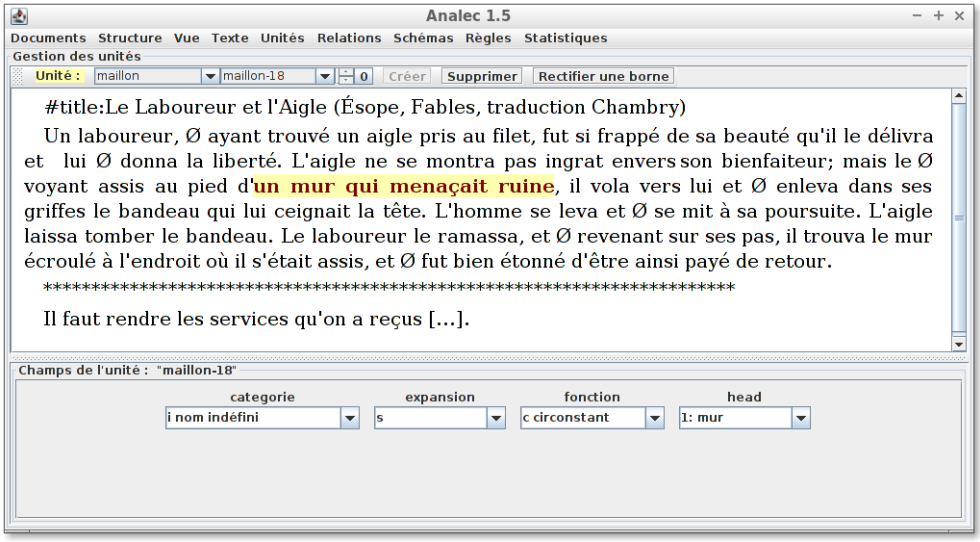

Fig. 4: Modification des propriétés dans Analec.

Reprenant la structure unité-relation-schéma de Glozz (Landragin, 2014) décrite plus haut, il se différencie de ce programme en ce qu'il offre des fonctions de visualisation des chaînes, des « représentations géométriques d'analyses factorielles de correspondance ou [de] rendu coloré de tests de khi-deux » (Landragin, 2014).

De plus, un « nouveau module pour le logiciel Analec », décrit par Landragin (2016) et « développé pour le projet ANR Democrat » a été conçu spécifiquement pour l'analyse des chaînes de référence. Il permet notamment :

- le calcul automatique des statistiques générales sur les chaînes (nombre moyen de maillons par chaîne, longueur moyenne des maillons, etc.),
- le filtrage des chaînes par nom du référent et par propriété,
- le découpage par paragraphe.

Ce module dispose aussi d'un calcul de bi-grammes et tri-grammes (déjà évoqué dans Landragin, 2014).

Néanmoins, ni à l'heure de la réalisation de nos annotations (mars-avril 2016), ni à celle de l'écriture de ces lignes (mai 2016), le module n'était publiquement disponible. Nous n'avons donc pas pu nous en servir.

Pour finir, il faut noter que si Analec enregistre les données sous son propre format binaire (donc lisible ni par un humain, ni avec un autre programme), il permet d'exporter et d'importer au format Glozz (*i.e.* XML).

4 Aperçu de nos scripts

Nous décrirons dans un instant les besoins que nous avons eus et qui, n'étant satisfaits ni par Glozz, ni par Analec, ni par aucun des autres logiciels d'annotation dont nous avons eu connaissance, nous ont conduit à programmer de nouveaux outils. Mais nous voudrions d'abord donner un rapidement aperçu de ces nouveaux outils.

Nous avons écrit deux jeux de scripts. Le premier est une interface graphique (voir figure 5) écrite en HTML et Javascript[20], et utilisable dans un navigateur Internet tel que Firefox ou Chromium/Google Chrome. Cela permet d'utiliser les puissantes fonctionnalités graphiques des navigateurs avec des coûts en termes d'effort de programmation et de ressources matérielles nécessaires relativement faibles (par comparaison, du moins, avec les coûts que de telles fonctionnalités demanderaient en Java, langage utilisé par Glozz et Analec). La limite la plus contraignante est l'impossibilité de créer des annotations qui se chevauchent, mais cela ne pose pas de problème puisque les expressions référentielles que nous annotons sont des syntagmes, qui, selon le modèle de la syntaxe générative, s'imbriquent les uns dans les autres et ne se chevauchent jamais. Les annotations imbriquées, comme dans (39) :

(39) [Le petit chat de [la voisine]$_j$]$_i$

sont parfaitement possibles (comme le montre la figure 5).

[20]HTML est un ensemble de balises qui permettent de mettre en forme les pages Internet, Javascript est un langage de script qui permet de rendre les pages Internet dynamiques.

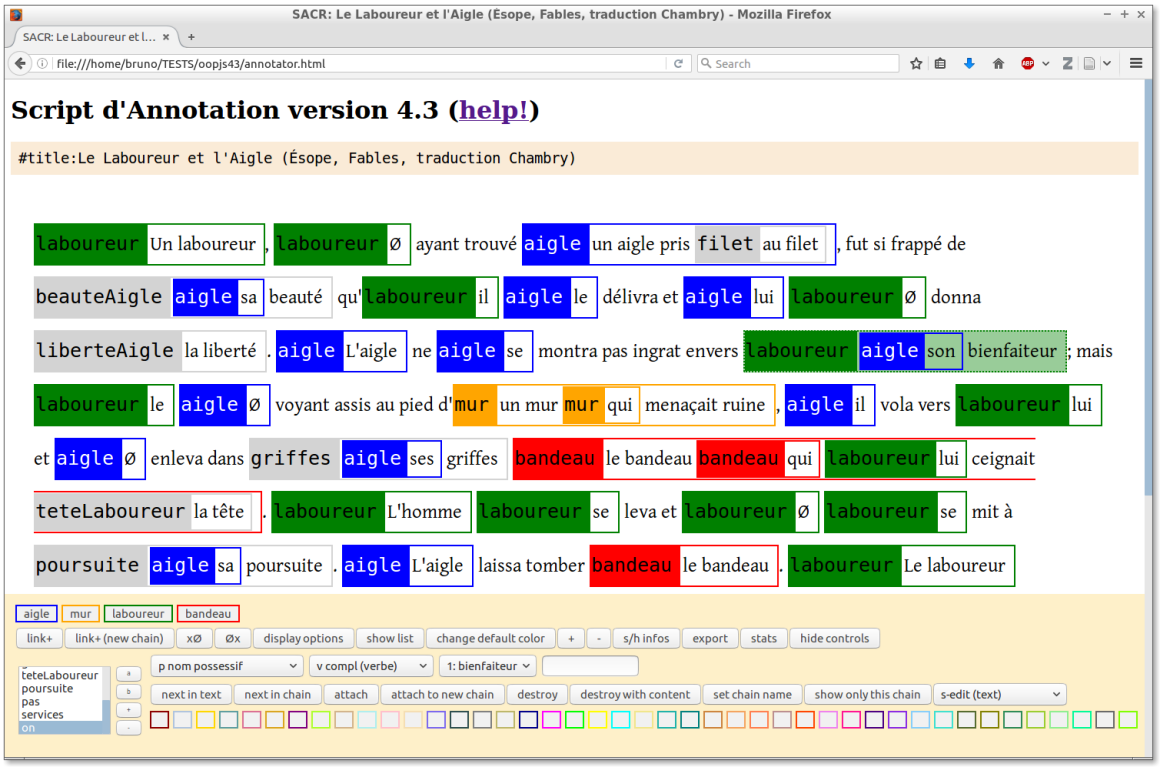

FIG. 5: Capture d'écran de l'interface.

Cette interface est destinée à l'annotation des textes, c'est-à-dire la délimitation des maillons et l'attribution, pour chaque maillon, des propriétés linguistiques (par exemple, la catégorie grammaticale, la fonction, les expansions). Elle contient également une fonction de recherche des maillons (selon le nom du référent ou les propriétés définies), et permet de visualiser l'ensemble des maillons sous forme de liste. C'est cette interface qui est décrite en détail dans l'annexe C : nous y renvoyons le lecteur pour les détails techniques.

L'autre jeu de scripts est écrit en Perl[21] (en programmation procédurale ou orientée objet[22] selon le script), et s'utilise en ligne de commande[23]. Il offre plusieurs fonctionnalités, décrites plus en détail dans les pages qui suivent :

- concordancier,
- calculs statistiques (nombre de chaînes et de maillons, distance intermaillonnaire, densité, etc.) et de fréquences absolues et relatives des propriétés linguistiques annotées, exportés au format CSV, lisibles par un tableur tel que Microsoft Excel ou LibreOffice Calc,
- modification, ajout, suppression et fusion d'annotations,

[21] Perl est un langage de script qui permet notamment de traiter automatiquement des données textuelles. Il s'exécute de façon autonome, c'est-à-dire qu'il n'a pas besoin d'un navigateur pour fonctionner.

[22] Il s'agit de deux techniques de programmation. La programmation orientée objet offre plus de souplesse et de possibilités, mais demande un effort plus important.

[23] Sous Linux. Il n'a pas été testé sous Windows, même s'il n'y a *a priori* pas de raison pour qu'il n'y fonctionne pas.

- conversion vers le format Glozz (et donc vers Analec, qui peut lire et écrire ce format[24]) et vice-versa,
- calcul automatiques des patrons[25] de chaînes les plus fréquents.

Ces deux jeux de scripts utilisent un format de fichier commun (décrit dans l'annexe C), qui un simple texte brut (c'est-à-dire sans formatage) avec des « annotations embarquées » (Habert, 2005, p. 32), c'est-à-dire intégrées au texte lui-même, un peu comme des annotations XML[26] :

```
{Chat:categorie="SN défini",fonction="sujet" Le petit chat
de {Voisine:categorie="SN défini",fonction="cdn" la voisine}}...
```

Alors que Glozz utilise des annotations déportées (*i.e.* dans un fichier séparé) et qu'Analec utilise un format binaire (*i.e.* lisible uniquement dans Analec), les annotations embarquées en texte brut ont l'avantage de pouvoir être lues par un humain avec n'importe quel éditeur de texte, et de pouvoir être immédiatement traitées par de petits scripts *ad hoc* ou des utilitaires Unix bien pratiques comme `grep` [27]. Certaines lectures et modifications peuvent donc être faites à moindre frais par ces moyens.

Nous avons développé ces scripts pour satisfaire au plus près nos besoins en termes d'annotation et d'analyse des données, besoins que ni Glozz ni Analec (ni les autres logiciels dont nous avons eu connaissance) ne pouvaient satisfaire. C'est pourquoi nous adopterons dans la suite de cette partie la démarche suivante : pour chaque besoin identifié, nous décrirons comment Glozz et Analec le satisfont (ou non), et comment nous avons résolu la difficulté à l'aide de nos propres scripts.

5 Besoins et solutions

Nous ne présenterons pas ici les problématiques générales liées à l'annotation de corpus (voir par exemple, Habert, Nazarenko et Salem, 1997 et Habert, 2005), et ne ferons qu'évoquer les problématiques liées à l'annotation des chaînes de référence et présentées par Landragin (2011, 2014) et Mélanie-Becquet et Landragin (2014). Par contre, nous nous concentrerons sur les besoins spécifiques à notre sujet, et qui n'ont pas été abordé, à notre connaissance, dans la littérature. Nous décrirons nos besoin en ce qui concerne l'annotation d'abord, puis en ce qui concerne l'analyse quantitative et qualitative.

[24] Les fichiers d'Analec sont binaires, c'est-à-dire qu'ils ne peuvent être lus que par ce programme, et c'est pourquoi nous avons été obligé, ou du moins il a été plus simple, de passer par le format Glozz, qui est en XML.

[25] Un patron est la séquence, régulière ou réitérée, des catégories grammaticales des maillons qui composent une chaîne. Voir ci-dessous, page 58, pour des exemples.

[26] Notre format est en fait un « XML dégradé », c'est-à-dire que nous avons supprimé les balises de fin et allégé la syntaxe des attributs. De la sorte, il est plus plus léger et plus facile à parser, et la conversion en XML reste très aisée.

[27] Il s'agit d'un outil qui permet, de façon très rapide et très efficace, d'extraire des informations d'un fichier texte à partir d'expressions rationnelles (ou régulières). Ces expressions sont des sortes de patrons qui permettent de récupérer les chaînes de caractères voulues dans un texte.

5.1 Besoins pour l'annotation

5.1.1 Besoins

Nous avons eu trois besoins majeurs en ce qui concerne l'annotation. D'abord, la possibilité de réutiliser le même schéma pour tous les textes du corpus. C'est le principe adopté par Glozz et Analec, et nous l'avons suivi. Ensuite, la rapidité et la facilité d'utilisation. Si Glozz est convivial et pratique, comme le rappelle Fort (2012), Analec est plus rudimentaire, notamment dans la délimitation des maillons (il travaille au niveau du caractère, alors que Glozz offre la possibilité de travailler, en surface au moins, au niveau du « mot ») et la sélection des propriétés : les clics de souris sont nombreux. Enfin, nous avons dû prévoir des propriétés particulières, notamment celle qui permet de déterminer la tête du syntagme.

5.1.2 Solutions

Nous nous étendrons surtout sur les deux derniers besoins, le premier n'offrant guère matière à discussion. Le but du jeu a été de trouver la façon d'annoter la plus rapide. En conséquence, nous avons d'abord choisi de travailler au niveau du token (que nous définissons ainsi : suite de lettres *ou* de chiffres, avec prise en compte des chiffres décimaux et des pourcentages (rattachés au nombre qui les précède), et avec possibilité de définir des tokens spéciaux (mots composés, symboles chimiques, lettres grecques, etc.) pour chaque texte si nécessaire) et non au niveau du caractère. En utilisant les simples liens hypertextes que tout navigateur reconnaît, il suffit de cliquer sur le premier mot du maillon et sur le dernier pour définir les bornes. Un bouton permet soit de créer une nouvelle chaîne avec le maillon : dans ce cas, un nom de chaîne, calculé à partir du contenu du maillon, est proposé. Un autre permet de rattacher le maillon à une chaîne existante. Si la chaîne a déjà plus de trois maillons, elle est accessible *via* un bouton coloré. (Tout cela est expliqué en détail et illustré dans l'annexe C.) Tout a été donc été pensé pour la plus grande ergonomie possible : dans la plupart des cas, il faut trois clics pour une annotation, deux au minimum (un maillon d'un seul token, par exemple un pronom, d'une chaîne déjà existante) et quatre au maximum (un maillon de plusieurs tokens d'une chaîne inexistante).

De même, la modification des bornes est très rapide : il suffit de sélectionner le premier et le dernier token du maillon, puis de cliquer sur le maillon à modifier : il faut donc deux (si le maillon n'a qu'un seul token) ou trois (si le maillon a plusieurs tokens) clics.

Nous venons d'évoquer des boutons pour rattacher un maillon à une chaîne : c'est que nous avons adopté un modèle uniforme de modélisation des chaînes : tous les maillons d'une même chaîne sont dans le même « sac », sans distinction autre que l'ordre d'apparition dans le texte. Il n'y a donc pas de description fine des relations (anaphorique ou coréférentielle, par exemple) entre les maillons. Cette possibilité existe (et est évoquée plus bas, et décrite dans le détail dans l'annexe C), mais nous n'avons pas eu le temps de l'utiliser dans ce travail. De plus, l'annotation des relations aurait pris beaucoup plus de temps.

Nous avons aussi essayé de rendre l'annotation des propriétés la plus rapide possible. En usant d'un avantage offert par les listes déroulantes des navigateurs, il est possible de définir pour

chaque valeur un code d'une lettre (« s » pour sujet, etc.). De plus, nous avons trouvé qu'annoter chaque propriété pour tous les maillons (par exemple annoter la fonction pour tous les maillons) est plus rapide qu'annoter toutes les propriétés pour chaque maillon (par exemple annoter la fonction, la catégorie grammaticale, et la tête syntaxique pour chaque maillon). En effet, la charge cognitive est bien moins lourde. Dans ce cas, l'annotation devient extrêmement rapide :

- un maillon est mis en relief, c'est-à-dire en couleur (et le texte défile automatiquement pour mettre le maillon vers le centre de l'écran au besoin),
- il suffit d'appuyer sur la touche correspondante à la valeur choisie (par exemple, si on annote la fonction, « s » correspondra au sujet, « v » au complément du verbe, etc.),
- automatiquement le maillon suivant est mis en relief (et se présente au centre de l'écran si besoin), et on recommence.

Il n'y a donc aucun clic de souris ni aucune touche qui est enfoncée inutilement : une frappe de touche correspond à une annotation ; on ne peut pas faire plus économique...

Nous avons aussi essayer de rendre l'interface attrayante et de faciliter le repérage des maillons et des chaînes, et nous nous sommes servi pour cela de la technologie offerte par les navigateurs Internet, qui mettent au service de l'utilisateur de nombreuses possibilités à peu de frais. Nos annotations sont en couleur, et chaque chaîne a automatiquement sa propre couleur. Les maillons imbriqués sont bien visibles puisque les cadres s'imbriquent les uns dans les autres (voir la figure 5), un peu comme dans Glozz (voir la figure 2), mais à la différence d'Analec qui ne permet pas de voir les maillons imbriqués de façon nette.

De plus, la structure textuelle est nettement visible, notamment par le découpage en paragraphe, mais aussi grâce aux inter-titres, qui, s'ils ont été marqués comme tels par des métadonnées, apparaissent en police de grande taille et en gras. La lecture et le travail s'en trouvent facilités.

Nous pensons que notre interface permet des annotations beaucoup plus rapides qu'Analec, et avec une présentation plus claire qui permet d'éviter les erreurs. C'est pourquoi nous avons utilisé notre script, plutôt que le programme de Democrat.

En ce qui concerne le dernier besoin, celui concernant les propriétés spécifiques, nous prendrons l'exemple de la tête syntaxique (voir page 80 pour la justification d'une telle propriété). Nous n'avons pas voulu entrer à chaque fois la position de la tête et son texte, c'est pourquoi cela est calculé automatiquement. Cela signifie que les valeurs que l'utilisateur peut choisir pour cette propriété changent pour chaque maillon, et elles sont mis à jour dès que les bornes du maillon changent. Cela permet une annotation extrêmement rapide pour cette propriété, puisqu'il suffit de frapper les touches « 0 » (si la tête est juste après la borne initiale, comme c'est le cas pour les pronoms) ou « 1 » (un SN avec un déterminant par exemple), plus rarement « 2 » (un SN avec un déterminant et un adjectif antéposé), etc.

5.2 Besoins pour l'analyse quantitative

5.2.1 Besoins

Le nombre de chaînes, le nombre de maillons dans le texte, le nombre moyen de maillons par chaîne, la distance intermaillonnaire, la densité (nombre de maillons par texte, par partie, par paragraphe), les coefficients normalisés de stabilité, etc. ne peuvent pas être calculés à la main dès lors que nous avons affaire à plusieurs milliers de maillons. Le calcul doit être automatique, et n'est proposé ni par Glozz, ni par Analec (le nouveau module d'Analec, décrit par Landragin (2016), propose un certain nombres de ces calculs, mais pas ceux qui nous introduirons dans le chapitre 6, comme le nombre d'îlots formées par une « chaîne partagée »).

Il en va de même pour le calcul automatique des fréquences, absolues et relatives, des classes de chaque propriété (par exemple, combien il y a de « SN définis » dans telle chaîne, tel paragraphe, telle partie, etc.).

De plus, il est intéressant de pouvoir diviser, ou au contraire regrouper, certaines propriétés. Ainsi, nous avons spécifié le type d'expansion (adjectif, nom, etc.), mais nous n'avons pas annoté explicitement la simple présence d'une expansion. De même, nous n'avons pas indiqué explicitement le nombre total d'expansions, ni le nombre total d'adjectifs, de compléments du nom, etc. Il faut donc pouvoir calculer ces propriétés « à la volée », entre le moment de l'annotation et le moment de l'interrogation des données. Il faut aussi pouvoir récupérer, par analyse morphologique de la tête du syntagme, le nombre voire le genre de l'expression (nous n'avons pas fait cela, par manque de temps, mais nous avons l'intention de le faire dans nos prochaines études).

De plus, puisque notre problématique touche à la différence, non seulement entre les différentes parties au sein d'un même article, mais aussi en fonction du type de partie (pour pouvoir avoir des statistiques sur toutes les introductions, par exemple, de tous les textes du corpus), il faut non seulement repérer les sauts de parties, mais aussi le type de partie auquel on a affaire. De plus, il a fallu couper les chaînes au niveau de chaque partie (par exemple, une chaîne peut s'étendre sur tout un texte, mais si l'on veut comparer les introductions, et seulement les introductions, de tous les textes, il faut couper la chaîne à la fin de l'introduction).

La deuxième étude exploratoire que nous proposons, elle, s'attache, entre autres, à faire la différence entre des chaînes qui n'apparaissent que dans un seul paragraphe et des chaînes qu'on retrouve dans plusieurs paragraphes. Il nous a fallu traiter ces dernières selon deux angles : à l'échelle du texte et à l'échelle du paragraphe. C'est pourquoi nous avons dû implémenter un module pour couper, ou non, au choix de l'utilisateur, les chaînes au niveau du paragraphe (mais pas toujours). La nouveau module pour Analec, présenté par Landragin (2016), semble proposer une telle fonction.

Notre but était également de proposer un classement des référents, c'est pourquoi le filtrage des chaînes était indispensable : il s'agit alors de demander au programme de ne calculer les statistiques et les fréquences uniquement pour les chaînes A, B et C, par exemple, et non pour toutes les autres. C'est une telle fonctionnalité qui nous a permis, par exemple, d'étudier les différences entre « chaînes uniques » et « chaînes partagées » (voir page 124).

Nous aurions aussi pu avoir besoin d'une fonction pour filtrer les textes, par exemple pour opposer les textes des sciences de la nature aux textes des sciences humaines. C'est un besoin que nous avons pris en considération, même si le faible nombre de textes annotés (cinq pour la première étude, quatre pour la deuxième) le rendait quelque peu superflu ; mais cela s'avèrera peut-être utile dans nos prochaines études.

Nous avons aussi pensé à pouvoir inclure ou exclure les inter-titres. En effet, la prise en compte ou non des inter-titres peut changer le format des chaînes. C'est une question que nous avons déjà évoquée quand nous décrivions la constitution de notre corpus (chapitre 2), et nous rappellerons que le comportement d'une chaîne peut changer si elle est initiée dans un titre (Jacques, 2005), même si nous n'avons pas eu le temps d'exploiter ces données dans les études exploratoires qui vont suivre.

Enfin, il nous a fallu un moyen de calculer automatiquement des patrons de chaînes. Un patron est la séquence des catégories grammaticales des maillons d'une chaîne. Par exemple, la chaîne :

(40) J'ai vu [un chat]$_i$. [Il]$_i$ était noir et [il]$_i$ est passé sous une échelle.

a pour patron :

(41) SN indéfini... Pronom personnel... Pronom personnel...

En observant un grand nombre de chaînes (pour peu qu'elles soient homogènes, notamment en ce qui concerne leur longueur), on peut voir émerger un ou plusieurs patrons d'ensemble. C'est ainsi que Schnedecker et Longo (2012) ont pu dégager trois patrons majoritaires dans les faits divers (en ne prenant en compte que les trois premiers maillons des chaînes) :

(42) a. SN défini... Pronom... Pronom...
b. SN indéfini... Pronom... Pronom...
c. SN défini... SN défini... SN défini...

Là aussi, un module de calcul automatique est indispensable.

5.2.2 Solutions

Aucun de ces besoins n'est satisfait par Glozz ou Analec (le nouveau module d'Analec en satisfait manifestement certains, mais nous n'y avons pas eu accès).

Un certain nombre de ces besoins requièrent l'insertion de métadonnées dans le texte original. Il s'agit :

- de la séparation entre les parties,
- de l'identification du type de partie,
- de l'identification des paragraphes qui sont des inter-titres (avec leur niveau hiérarchique).

Les métadonnées sont ajoutées dans le fichier du texte, chacune sur une ligne commençant par le symbole `#`. Par exemple (l'annexe C décrit l'implémentation de ces métadonnées plus en détail) :

```
#part-type:introduction

#part-heading:level=1

Histoire du petit chat

Cela fait bien longtemps que ma voisine a adopté son chat...
```

Les classes de référents, elles, sont identifiées par le nom donné à la chaîne. Le nom commence par une majuscule : toute minuscule qui précède est considérée comme le code d'appartenance à un groupe. Par exemple, nous avons choisis le code « e » pour les référents qui sont des ensembles. Nous avons donc des chaînes `eEtudiants` , `eParticipants` , etc. Un même référent peut appartenir à plusieurs groupes (en mettant plusieurs codes avant la majuscule), même si nous ne nous sommes pas servi de cette fonctionnalité (nos classes ont été pensées comme exclusives).

Le *parser*[28], écrit en Perl, accepte des informations de filtrage au niveau des chaînes et/ou des textes, et des options qui permettent par exemple d'inclure ou d'exclure les titres, mais aussi de supprimer les chaînes de moins de trois maillons (ce qui arrive lors du découpage en paragraphes, par exemple une chaîne de trois maillons, dont les maillons apparaissent chacun dans un paragraphe différent, n'est plus une chaîne une fois le découpage en paragraphe effectué).

De plus, une fonctionnalité permet d'ajouter (ou d'éliminer) des propriétés, calculées à partir de celles déjà présentes.

Nous avons écrit le script en programmation orientée objet, ce qui permet de traiter chaque chaîne comme une entitée séparée appartenant à la fois à un paragraphe, une partie, un texte et au corpus. Cela doit être pris en considération pour le calcul des moyennes. En effet, la moyenne des moyennes de n groupes d'éléments n'est pas égale à la moyenne de l'ensemble des éléments pris individuellement. Par exemple, imaginons qu'un élève a eu 4 et 8 en mathématiques, et 10, 12, et 14 en français. Sa moyenne générale pourra être calculée de deux façons :

$$\frac{\frac{4+8}{2} + \frac{10+12+14}{3}}{2} = 9.5$$

ou

$$\frac{6+8+10+12+14}{5} = 10$$

Nos moyennes sont des moyennes portant sur l'ensemble des chaînes, et non des moyennes de moyennes. Ainsi, la distance intermaillonnaire moyenne des chaînes du corpus n'est pas la moyenne des moyennes des distances au niveau du texte, mais bien la moyenne de toutes distances intermaillonnaires des chaînes du corpus, calculée directement, sans passer par le

[28]Ou « analyseur syntaxique » en français. Il s'agit d'un programme qui transforme un texte lisible par un humain en une représentation compréhensible par un ordinateur.

niveau du texte. Il en va de même pour les parties : les moyennes au niveau de la partie sont calculées directement, sans passer par la moyenne des paragraphes.

Au final, le script Perl que nous avons écrit donne des fichiers CSV, qui sont lisibles par un tableur comme Microsoft Excel ou LibreOffice Calc. Pour chaque sélection de référents (c'est-à-dire pour l'ensemble des référents, mais aussi pour chaque groupe de référents, ou bien pour chaque filtrage de référents et/ou de textes : les combinaisons sont nombreuses), nous avons quatre fichiers :

- l'un contient les données statistiques globales (nombre de tokens, nombre de paragraphes, nombre de chaînes, de maillons, densité, distance intermaillonaire moyenne, nombre de paragraphes couverts par une chaîne, nombre d'îlots, etc., etc.) ; cela est calculé pour l'ensemble des référents (soit dans tout le corpus, ou dans le groupe, ou dans le filtre) dans le corpus entier, dans chaque texte, dans chaque partie de chaque texte, dans chaque paragraphe de chaque partie de chaque texte, et enfin pour chaque type de parties (toutes les introductions, méthodologies, résultats, discussions et conclusions du corpus),
- un deuxième fichier contient les mêmes statistiques, mais calculées pour chaque chaîne, dans tout le corpus, dans chaque texte, chaque partie, chaque paragraphe et chaque type de partie. Cette disposition permet de facilement comparer les différentes chaînes à l'intérieur d'un paragraphe, d'une partie, etc.,
- un troisième fichier contient toutes les fréquences (absolues et relatives) de toutes les propriétés ; cela pour l'ensemble des référents (corpus, groupe ou filtre) de tout le corpus, ainsi que pour chaque texte, etc. (comme pour le premier fichier),
- un dernier contient les mêmes fréquences, mais calculées cette fois pour chaque chaîne, dans tout le corpus, dans chaque texte, etc. (comme pour le deuxième fichier).

Le script d'analyse des données provoque donc une avalanche de chiffres. Il ne s'agit évidemment pas de tout lire. Comme nous l'avons expliqué plus haut, nous avons adopté une approche descendante : partant des données du corpus global, nous avons cherché ensuite à analyser les oppositions entre textes (sciences naturelles et sciences humaines), entre parties et entre types de référents (abstraits, ensembles, etc.). Nous avons donc utilisé seulement certains de ces fichiers. Les autres nous ont servi à l'investigation de détail : chaque fois que nous avions repéré un phénomène intéressant au niveau global, nous avons tenté de l'expliquer en resserrant le niveau d'analyse : passant du corpus au texte, puis à la partie, au paragraphe et enfin à la chaîne. Cela nous a permis de faire des analyses fines (parfois il ne s'agissait que d'erreurs d'annotation), et la présence de ces fichiers CSV en grand nombre nous a été très utile. En dernière analyse, nous pouvions retourner au texte. Par exemple, si le fichier CSV nous indiquait un phénomène sur une chaîne A dans le paragraphe 36 du texte 2, nous pouvions tout de suite aller voir les annotations.

De plus, comme nous avons utilisé un tableur (LibreOffice Calc dans notre cas) pour lire ces fichiers CSV, nous avons pu profiter de la puissance du logiciel pour calculer d'autres valeurs (moyennes, sommes, coefficients de corrélation, etc.) ; et surtout pour visualiser rapidement les données en créant des graphiques « à la volée ».

Il serait même possible d'utiliser ces fichiers avec d'autres programmes, spécialisés en statistiques (nous pensons à R Studio). Nous ne l'avons pas fait, d'abord par manque de temps,

ensuite parce que nous avions trop peu de données pour que cela en vaille véritablement la peine.

Il nous reste enfin à parler des patrons de chaînes. Un script séparé calcule l'ensemble des patrons et les classe par fréquence, en affichant, selon le choix de l'utilisateur, les groupes les plus représentés dans chaque branche. D'autres options permettent de spécifier la profondeur, ou bien de définir un seuil au-dessous duquel les branches ne sont plus affichées. Le tout se veut une sorte d'« arbre de probabilités », bien que, puisque nous avons si peu de données, il s'agisse plutôt d'un « arbre de proportions » (comme dirait Swales (1990, p. 170), notre étude est « largely restricted to an exploratory rather than hypothesis-testing stage »).

La figure 6 illustre un tel arbre (il s'agit bien sûr d'une mise en forme, la sortie brute du script contient bien plus de données).

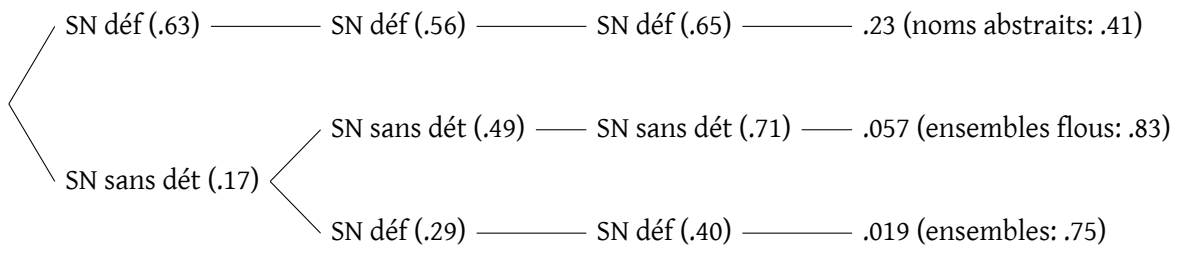

Fig. 6 : Exemple d'un arbre qui indique les patrons les plus courants (repris de la figure 8, page 121)

Comment lire un tel arbre ? Le premier niveau montre le premier maillon : dans 63 % des cas, il s'agit d'un SN défini, dans 17 % des cas, d'un SN sans déterminant. Les autres valeurs ne sont pas réprésentées ici, car nous ne les avons pas jugées significatives : nous avons fixé le seuil à 10 %. Le deuxième niveau montre, logiquement, le deuxième maillon. Si le premier maillon est un SN défini, alors dans 56 % des cas (les autres cas, jugés non significatifs, non pas été représentés), le deuxième maillon est aussi un SN défini. Mais si le premier maillon est un SN sans déterminant, alors le deuxième maillon est dans 49 % des cas un SN sans déterminant et dans 29 % un SN défini. La démarche est la même pour le troisième maillon.

Enfin, chaque branche se termine par la proportion de la branche entière : 23 % de *toutes* les chaînes ont le patron :

(43) SN défini... SN défini... SN défini...

alors que 5,7 % ont le patron :

(44) SN sans déterminant... SN sans déterminant... SN sans déterminant...

et 1,9 % ont le patron :

(45) SN sans déterminant... SN déf... SN sans déterminant...

Les autres patrons possibles sont négligeables (le patron (45) présenté ici est lui-même négligeable par rapport au patron (43)).

Tout au bout de chaque branche apparaît le type de référent (voir le chapitre 5) le plus représenté dans la branche. Par exemple, 41 % des chaînes qui ont le patron (43) sont des noms abstraits (voir page 101 pour ce que nous entendons par là).

Si nous avions plus de données, nous pourrions faire des prédictions, et dire par exemple que si le premier maillon est un SN défini, alors il y a plus d'une chance sur deux pour que le deuxième maillon soit aussi un SN défini. Et si nous avons deux SN définis, alors il y a 65 % de chances pour que le troisième maillon soit de la même catégorie grammaticale. Et si c'est le cas, alors nous avons 41 % de chances pour que ce soit un nom abstrait.

On peut aussi envisager, même si nous ne l'avons pas fait, d'étudier de la même manière la fonction grammaticale. Cela permettrait de caractériser le premier maillon (ou le deuxième, etc.), par exemple s'il est le plus souvent sujet, objet, circonstant, etc.

5.3 Besoins pour l'analyse qualitative

Pour des analyses plus qualitatives, nous avons eu besoin de voir les annotations dans le texte lui-même. Les deux jeux de scripts nous y ont aidé.

L'interface graphique, d'abord, permet de faire des recherches, simples mais suffisantes dans la majorité des cas. Il est possible de n'afficher que certaines chaînes (au choix de l'utilisateur), et/ou de n'afficher que les maillons qui répondent à certains critères relatifs aux propriétés. Par exemple, il est possible d'afficher tous les maillons des chaînes A et B (mais pas les autres) qui sont sujets mais ne sont ni des SN définis ni des pronoms démonstratifs.

L'interface graphique permet aussi d'afficher la liste de tous les maillons de toutes les chaînes, par ordre d'apparition, et, d'un simple clic, de les retrouver dans le texte lui-même.

Enfin, cette interface permet de visualiser la répartition des chaînes dans le texte. Celui-ci est réduit à une succession de point (chaque point représente un token), et ces points, s'ils représentent des maillons, sont soulignés de la couleur de chaque référent. Les mêmes possibilités de recherche sont offertes que dans le texte (recherche sur les propriétés).

L'ensemble des ces fonctionnalités sont décrites et illustrées par des copies d'écran dans l'annexe C.

Du côté des scripts d'analyse, nous avons intégré un concordancier, dont les modes d'affichage sont définissables *via* un script Perl. Le concordancier permet d'afficher les occurrences pour tout le corpus, ou bien par texte, par partie ou par paragraphe ; puis, en second niveau de tri, par référent, type de référent, ou par propriété. Il est possible, à tous les niveaux, de définir des conditions (par exemple tous les pronoms sujets), et d'afficher les résultats positifs mais aussi, selon le choix de l'utilisateur, les résultats négatifs (ce qui permet des comparaisons). Les fréquences sont calculées automatiquement (résultats positifs/négatifs).

Il est possible de définir la taille du contexte, ainsi que de l'arrêter aux signes de ponctuation. Cela nous semble plus efficace pour trouver les mots les plus fréquents dans le contexte gauche et droit, qui sont automatiquement affichés à la fin de chaque section d'affichage (texte, partie, paragraphe, propriété).

```
bruno ~/TESTS/read_annotations2
File  Edit  Tabs  Help
*****************************************************************************
  TEXT: text: annotated_texts/test_aesop_20160531-151546
*****************************************************************************


  PART: part: 1 (no type)

===== annotated_texts/test_aesop_20160531-151546:aigle =====
----- categorie: d nom défini -----
- L'aigle ne se montra
- L'aigle laissa tomber le
----- categorie: e déterminant possessif -----
- si frappé de sa beauté qu'il
- pas ingrat envers son bienfaiteur
- Ø enleva dans ses griffes le bandeau
- se mit à sa poursuite
----- categorie: f pronom réfléchi -----
- L'aigle ne se montra pas ingrat
----- categorie: i nom indéfini -----
- Ø ayant trouvé un aigle pris au filet
----- categorie: s pronom personnel -----
- beauté qu'il le délivra et  lui
- le délivra et  lui Ø donna la
- il vola vers lui
----- categorie: z pronom zéro -----
- mais le Ø voyant assis au
- vers lui et Ø enleva dans ses
TOTAL: accepted: 13, rejected: 0
===== annotated_texts/test_aesop_20160531-151546:bandeau =====
----- categorie: d nom défini -----
- dans ses griffes le bandeau qui lui ceignait la tête
- aigle laissa tomber le bandeau
----- categorie: r pronom relatif -----
```

FIG. 7: Une vue du concordancier en ligne de commande.

Ce concordancier, en ligne de commande mais tout en couleur (voir figure 7), permet donc de faire des recherches de détails et d'afficher les contextes dont on a vraiment besoin. Cependant, il n'est pas interactif.

5.4 « Besoins » non exploités

Nous voudrions terminer ce tour d'horizon par des besoins que nous avons ressentis dès le début de l'annotation, et que nous avons donc cherché à satisfaire par nos scripts, mais que nous n'avons pas exploité dans les deux études exploratoires que nous présenterons plus tard, faute de temps.

Nous avons ainsi créé un système qui permet de mettre en relation deux maillons. C'est le principe des « relations » de Glozz et Analec, mais de façon bien plus rudimentaire (le maillon d'ancrage est simplement une propriété du maillon cible, et les relations ne sont pas des objets en tant que tels). Nous avions envisagé ce système pour définir des relations entre chaînes, par exemple des relations d'anaphore associative (comme ici entre i et j).

(46) Nous arrivâmes dans [un village]$_i$. [L'église]$_j$ était en hauteur.

Il est aussi possible de se servir de cette fonctionnalité pour annoter les relations entre les maillons d'une même chaîne (par exemple cataphore, ou anaphore *vs* coréférence, etc.).

Par ailleurs, nous avions prévu de mettre des annotations sur les chaînes dans les métadonnées, comme c'est déjà le cas, par exemple, pour la couleur de la chaîne (même si cette information est, pour l'heure, triviale).

Nous avons également défini des chaînes spéciales (dont le nom commence par le caractère _), pour annoter, par exemple des marqueurs discursifs, et observer les maillons par rapport à ces marqueurs. Charolles, 1988, en effet, présente différents plans d'organisation textuelle : périodes, chaînes, portées et séquences. Ces différents plans d'organisation du discours ne sont pas autonomes (p. 11), ils sont au contraire « en interaction permanente ». Par exemple, « la constitution des chaînes est gouvernée par la configuation des périodes dans lesquelles entrent les expressions référentielles » (p. 11). Or certains de ces plans sont introduits par des « marqueurs discursifs », qui sont des expressions qui signale la structuration du discours (l'exemple le plus évident est celui des cadratifs temporels (Ho-Dac, 2005)). Ho-Dac (2005) étudie la relation entre ces marqueurs et les chaînes de référence, notamment dans la perspective d'une analyse quantitative.

C'est pourquoi nous avons programmé l'interface de manière à ce que les propriétés puissent être différentes en fonction de la chaîne (le codage se fait en fonction du nom de la chaîne). Un marqueur discursif n'aura en effet pas les mêmes propriétés qu'une expression référentielle. Cela est également possible, et bien plus efficace, dans Glozz et Analec.

6 Compatibilités et conversions depuis et vers Glozz et Analec

Nous avons programmé nos scripts pour répondre à des besoins précis. Analec répond à des besoins un peu différents, et permet des calculs statistiques que nous n'avons pas pu programmer. Nous avons donc créé des scripts de conversion entre notre propre format et le format Glozz, lisible par Analec. Cela nous permet de naviguer entre nos propres scripts (interface et analyse), Glozz (qui est trop gourmand en ressources pour que nous puissions vraiment l'utiliser) et Analec. De plus, nos scripts de conversion permettent de calculer de nouvelles propriétés.

Nous suivons en cela les recommandations de Landragin, 2014 :

> En général, ce n'est pas un logiciel mais la combinaison de plusieurs logiciels qui permettra de procéder à un maximum de calculs et de répondre à un maximum de questions. Comme nous l'avons évoqué [plus haut] et montré ensuite avec des allers et retours constants entre GLOZZ, ANALEC et Excel, la solution ne se trouve que rarement dans l'exploitation d'un logiciel unique, mais plutôt dans la mise en oeuvre d'une cascade de logiciels.

Même si nous avons développé ces scripts de conversion pour profiter des avantages d'Analec, nous n'avons pas eu le temps de nous en servir pour les deux études que nous présentons ci-dessous. C'est donc dans des études ultérieures que nous utiliserons les modules statistiques d'Analec.

7 Conclusion

Après avoir décrit notre processus d'annotation, nous avons présenté deux logiciels, Glozz et Analec, qui sont utilisés pour l'annotation des chaînes de référence. Néanmoins, des besoins

spécifiques à notre problématique nous ont amené à développer de nouveaux outils (une interface en HTML et Javascript, et des scripts pour l'analyse des données en Perl).

Ces outils ne seraient que peu de valeur sans propriétés linguistiques à annoter : c'est ce que nous allons maintenant expliciter.

Chapitre 4

Justification du schéma d'annotation

Nous présenterons dans cette partie le schéma d'annotation que nous avons établi pour annoter notre corpus. Nous évoquerons les principes que nous avons suivis pour son élaboration puis nous justifierons les choix que nous avons dû faire. Ce faisant, nous constituerons un « petit guide d'annotation » qui pourra servir à la fois de notice explicative des décisions que nous avons prises et de référentiel pour nos futures annotations.

Cependant, nous ne discuterons ici que d'aspects grammaticaux ; les problèmes référentiels, et les choix que nous avons faits à leur sujet, sont décrits et analysés dans le chapitre 5 (« Une première étude exploratoire : annotation d'une sélection de référents saillants »).

Il faut noter que l'élaboration de ce schéma a été un processus itératif. Nous sommes parti, pour les catégories grammaticales, du schéma utilisé par Schnedecker (2014) : c'est avec lui que nous avons commencé l'annotation de notre corpus. Mais nous avons dû l'adapter à la spécificité de nos textes et de notre méthodologie. Par exemple, nous avons ajouté une « fonction » « titre et parenthèse » puisque nous avons un nombre significatif de maillons dans des titres et des parenthèses : employés hors phrase, leur fonction ne correspond à aucune des catégories usuelles (sujet, complément, etc.). Par ailleurs, nous avons séparé l'annotation de l'expansion de celle de la catégorie grammaticale (au lieu de « SN défini expansé », nous avons deux propriétés : « SN défini » et « expansé »), car cela a facilité la programmation des scripts pour l'analyse quantitative.

1 Principes et compromis

Quatre principes nous ont guidé dans l'élaboration du schéma d'annotation. Nous avons d'abord cherché une caractérisation linguistique suffisante pour permettre une analyse linguistique fine. C'est ainsi, par exemple, que nous ne nous sommes pas contenté d'indiquer simplement si un GN était expansé, mais nous avons noté le type et le nombre d'expansions.

Cependant, et c'est là le deuxième principe, il nous a fallu limiter le nombre de classes dans chaque propriété annotée, afin d'avoir assez de données dans chacune d'elle, malgré le nombre réduit des textes de notre corpus, pour les deux études exploratoires qui seront présentées dans la suite. Cela nous a contraint, par exemple, à ne pas multiplier les fonctions, et à réunir sous l'appellation « complément du verbe » les compléments d'objet directs et indirects, ou encore à ne pas détailler les différentes valeurs sémantiques ou formelles des circonstants.

C'est par respect de ce principe, également, que nous avons considéré que les déterminants possessifs avaient la fonction de complément du nom (dans l'optique où « son chat » serait équivalent à « le chat de lui »), alors que, d'un point de vue strictement linguistique, ils n'ont pas de fonction. Mais créer une classe supplémentaire dans la propriété « fonction » nous aurait amené à créer une classe assez peu fournie, et en plus redondante avec la catégorie grammaticale puisque seuls les déterminants possessifs seraient entrés dans cette classe.

Limiter le nombre de classes oblige néanmoins à veiller à ce qu'elles restent homogènes : c'est là notre troisième principe. Si nous avons considéré que les déterminants possessifs avaient la fonction de complément du nom, nous avons tenu à les séparer des autres expressions référentielles au niveau de la catégorie grammaticale, et nous ne les avons pas inclus dans les pronoms personnels : cela aurait rendu la classe trop hétérogène, puisque le déterminant possessif est, comme son nom l'indique, un déterminant et non un SN, et aurait pu compromettre la pertinence des analyses quantitatives. La difficulté, ici, est de trouver un compromis entre les deuxième et troisième principes.

Enfin, nous avons cherché à rendre nos annotations compatibles avec le projet Democrat, afin de pouvoir les convertir vers le schéma de Democrat[29], même s'il s'avère que notre corpus ne pourra finalement pas faire partie de celui du projet[30]. Par exemple, nous avons choisi d'annoter les « sujets zéro » par un symbole particulier (« ø »), mais en veillant à toujours à la mettre avant le verbe. Cela nous permet en effet de bien les repérer dans les concordanciers et, de plus, c'est une condition *sine qua non* pour le calcul automatique du coefficient de stabilité formelle que nous introduirons un peu plus tard (page 82). Cependant, comme Democrat requiert d'annoter les sujets zéro en marquant le verbe (voir Landragin, 2011, p. 73 pour la justification), et non en insérant un symbole supplémentaire, la conversion peut se faire extrêmement simplement, puisqu'il suffit de transférer le balisage du symbole « ø » au mot suivant.

De même, nous avons annoté les pronoms réfléchis, sans les mettre dans la classe des pronoms personnels : cela aurait pu se justifier et aurait permis de limiter le nombre de classes, mais n'aurait pas été compatible avec l'annotation de Democrat, qui rejette des pronoms réfléchis. Nous avons, ici encore, cherché le meilleur compromis.

Nous allons maintenant décrire comment nous avons sélectionné les unités linguistiques à annoter, et comment nous les avons bornées.

[29] D'après les informations dont nous disposons au printemps 2016.
[30] Pour des questions techniques de droits d'auteur.

2 Délimitation des maillons

Nous commencerons par caractériser les expressions référentielles que nous avons annotées.

2.1 Éléments zéro

2.1.1 Sujets zéro

Il y a des « éléments non-référentiels qui participent… à la coréférence, notamment certains sujets zéro » (Landragin, 2011, p. 63). Les « sujets zéro » correspondent à ce que la GMF, par exemple, appelle « l'ellipse du sujet ». Outre la langue orale et les dialogues (notamment les questions-réponses), qui ne nous concernent pas, la GMF (p. 249) note que les seuls contextes où on peut trouver une telle ellipse sont :

- la phrase impérative,
- la mise en facteur commun d'un même sujet dans une coordination (voir aussi GMF, pp. 876–7),
- le sujet coréférentiel au sujet d'un autre verbe, notamment dans le cas d'une infinitive.

Nous avons marqué les deux premiers cas en insérant un symbole « ø », placé juste avec le verbe, ce qui permet la compatibilité avec l'annotation de Democrat, comme nous l'avons indiqué plus haut. L'impératif ne se rencontre guère qu'à la première personne dans les textes de notre corpus :

(47) Ø notons les différences… (T2)

Il faut cependant noter que l'absence du sujet est, dans le cas de l'impératif, inhérente au système de la langue : il n'est jamais présent, et ne peut pas alterner avec un pronom clitique. Cependant, nous pensons que ne pas marquer ce sujet reviendrait à ignorer une forte présence référentielle, et à gommer la différence fondamentale entre, par exemple, « Note ! », « Notons ! » et « Notez ! ».

La mise en facteur commun est beaucoup plus fréquente :

(48) Ils s'y intéressent sans doute et Ø adoptent peut être la vision de leur établissement (T2)

Nous n'avons pas annoté les sujets zéro d'infinitifs ou de participes, suivant en cela le projet Democrat, et bien que certaines études l'ait fait (Landragin, 2011, p. 73, Mélanie-Becquet et Landragin, 2014, p. 118).

2.1.2 Compléments zéro

Nous n'avons pas annoté, non plus, l'ellipse du complément. Celle-ci peut avoir lieu lorsque l'objet est « contextuellement restituable » (GMF, p. 396), comme dans :

(49) Je sais (ce que vous venez de me dire). (GMF)
(50) Regarde (ce que je te montre). (GMF)

ou bien quand « l'absence de réalisation lexicale de l'objet permet d'identifier le procès verbal en lui-même, sans autre spécification » (GMF, p. 396), comme dans :

(51) Je mange.

Le deuxième cas ne peut pas être anaphorique : nous n'avons donc pas à nous en occuper. Quant au premier cas, s'il est fréquent dans les recettes de cuisine (Schnedecker, 2014), semble insignifiant dans notre corpus.

2.1.3 Des expressions référentielles ?

Alors que certaines études annotent les sujets zéros comme nous l'avons fait, voire les compléments zéro (par exemple Schnedecker, 2014), d'autres (par exemple Longo et Todirascu, 2014, p. 85) choisissent de ne pas annoter ces expressions qui ne sont pas explicites. D'un côté, il peut sembler important de les prendre en compte, car « l'absence matérielle de sujet se compense par la conscience d'une présence subjectale implicite » comme l'indique Benayoun (2003, p. 173), qui étudie le sujet zéro dans les recettes de cuisine, le journal intime et les disdascalies, et pointe qu'« il y a sujet dès lors qu'il y a prédication ». Ainsi, dit-il, « la place du sujet est toujours matérialisée ou matérialisable, qu'il soit instancié ou non... Ce n'est pas parce que l'on ne voit ni n'entend le sujet, que sa référence n'est pas posée » (p. 174). C'est également ce que pense Landragin (2011, p. 72) : « Ce n'est pas parce qu'un tel élément n'est pas exprimé qu'il n'est pas saillant à la fois du point de vue du locuteur que du lecteur ». Et pourtant il choisit de faire la différence entre « maillon faible » (sujets zéro et autres marques implicites) et « maillon fort » (expressions explicites). C'est parce que « ce type de phénomène est toujours un obstacle aux méthodes d'annotation dans la mesure où l'on souhaite repérer et annoter du vide » (p. 72), ce qui fait écho à Lemaréchal (1997), cité par Neveu (2004, p. 305) :

> D'un point de vue épistémologique, poser des marques ø ou des constituants ø est suspect, puisque cela revient à poser un segment, constituant ou marque segmentale, dont le signifiant est représenté précisément par une absence de segment, donc à poser des segments fictifs... On est vite contronté à une difficulté supplémentaire : l'impossibilité fondamentale qu'il y a à catégoriser... de tels segments « qui n'existent pas ».

Nous avons annoté ces « segments fictifs » comme de vraies expressions référentielles, mais c'est un point qui mériterait d'être discuté de manière plus approfondie.

2.2 Pronoms relatifs

2.2.1 Pronoms introduisant des déterminatives ou des appositives

On distingue traditionnellement deux types de relatives. Les relatives déterminatives, d'abord, ont « le comportement d'un modifieur. L'absence de la relative entraîne un changement d'interprétation, telle que l'objet désigné par le groupe nominal change d'extension » (Delaveau, 2001, p. 95) :

(52) Le roman que je termine est très intéressant.

La relative est ici « intégrée au SN » (Charolles, 2007).

Les relatives appositives (ou explicatives), ensuite : « [leur] absence ne change pas l'interprétation du groupe nominal, qui garde la même référence, qu'elle soit présente ou non » (Delaveau, 2001, p. 96) :

(53) Ce roman, que je viens de finir, me plaît beaucoup.

Ici, « le groupe nominal doit pouvoir être interprété sans la contribution de la relative à la détermination de sa référence » (p. 96), si bien que « la relative n'est pas intégrée..., n'est pas nécessaire et elle peut être détachée de la matrice par un signe de ponctuation » (Charolles, 2007).

On peut, dans chaque cas, se demander si le pronom relatif réfère. En effet, dans le cas des déterminatives, puisque l'antécédent lui-même n'a pas de référence propre (car il a besoin de la relative pour que la référence puisse être calculée), on peut douter que le relatif en ait une. C'est en tout cas ce que pense Le Goffic (1993, p. 47) : « Dire que le relatif est anaphorique de son antécédent n'est pas totalement exact, dans la mesure où il contribue très fréquemment à le déterminer. »

Le deuxième cas semble poser moins de problèmes : il s'agit de l'expansion d'un nom dont la référence est bien établie en elle-même. On pourrait donc penser que le pronom relatif, puisqu'il est une anaphore de l'antécédent, est une expression référentielle comme n'importe quel autre pronom. Mais Charolles (2007) semble rejeter cette idée. Pour lui, en effet,

> le pronom relatif fonctionne d'une façon très différente du pronom de troisième personne et c'est cela qui explique que, même dans les emplois les moins intégrés syntaxiquement..., il n'a pas les mêmes pouvoirs référentiels que celui-ci... Le pronom relatif est une forme liée mémoriellement, il ne réactualise pas, ne réinitialise pas, le référent de son antécédent comme le fait le pronom [personnel]... À proprement parler, *le relatif ne réfère pas*, il se contente de maintenir son antécédent dans la mémoire des auditeurs/lecteurs (nous soulignons)

au contraire des pronoms personnels et démonstratifs, « formes pleinement référentielles, [qui] remettent en avant leur référent ».

Ainsi, d'après ce que nous comprenons, puisque le relatif « s'accroch[e] à un antécédent (SN plein ou pronom disjoint) qu'il recrute par proximité » et « ne réfère pas », Charolles ne considèrerait pas les pronoms relatifs comme des expressions à annoter dans une chaîne de référence.

Cependant, certains relatifs au moins nous semble pouvoir accéder pleinement au statut d'expression référentielle, comme les relatifs qui apparaissent après une préposition (*par lequel...*). De plus, à notre connaissance, aucun des auteurs qui ont annoté des chaînes de référence n'ont rejeté les pronoms relatifs d'un bloc. Tout au plus y a-t-il discussion sur l'inclusion des pronoms qui introduisent une relative déterminative. Le projet Democrat va plus loin encore, et demande d'inclure l'ensemble des relatifs. C'est ce choix que nous avons prudemment suivi.

2.2.2 Pronoms introduisant d'autres relatives

Il existe d'autres types de relatives ; elles ne sont cependant pas toutes évoquées par l'ensemble des auteurs.

Delaveau (2001, p. 97), par exemple, cite les relatives attributives :

(54) J'ai vu Paul qui entrait.

Introduites par un verbe de perception et décrivant une action simultanée à l'acte de perception (Leeman, 2002), nous n'en avons pas rencontré dans notre corpus.

Charolles (2007), lui, évoque les relatives « dites narratives », où le relatif commute « sans difficultés avec le pronom de troisième personne » pour former deux phrases :

(55) a. Louise fit signe à un/au garçon, qui lui apporta l'addition.
b. Louise fit signe à un/au garçon. Il lui apporta aussitôt l'addition.

La GMF, quant à elle, évoque les relatives périphrastiques, qui « constituent formellement l'expansion d'un pronom démonstratif... de manière à former avec lui l'équivalent d'un GN ». Elles ont « un statut intermédiaire entre celui des relatives adjectives... et des substantives proprement dites » (p. 814).

Néanmoins dans un souci d'homogénéité et pour éviter de prendre trop parti dans un cadre théorique ou dans un autre, nous avons annoté tous les pronoms relatifs, quelle que soit la relative dans laquelle ils apparaissent. Nous suivons en cela aussi le projet Democrat.

2.3 Noms quantifiants

Les noms quantifiants (*tas*, *nombre*, *ensemble*), entrent dans la composition de « déterminants composés » (GMF, p. 350). En effet, ils commutent avec des déterminants simples (*un tas de problèmes vs quelques problèmes*) et ne peuvent être repris par une anaphore (*Jean a un tas de problèmes. * Ce tas l'empêche de dormir*).

Nous les avons donc inclus dans le maillon de la tête du syntagme :

(56) [l'ensemble des Grandes Causses]$_i$ (T1)

2.4 Attributs

La fonction d'attribut est, en général, d'assigner une qualité ou un état au sujet (GMF, p. 419), il n'y a donc, *a priori*, pas lieu de les annoter :

(57) Le chat est un animal domestique.

Pourtant, il existe des cas où on pourrait hésiter. Le premier est analysé par Kleiber (1981, pp. 114 *sq.*) :

(58) a. Paris est la capitale de la France. (Kleiber)
b. La capitale de la France est Paris. (Kleiber)

Ce couple de phrases montre qu'il peut y avoir permutation entre le sujet et l'attribut, et on peut se demander si, dès lors que l'un et l'autre peuvent être, indifféremment, sujet, il ne faudrait tout de même pas de les considérer comme expression référentielle à annoter. En fait, il existe une réelle différence entre les deux expressions (*Paris* et *capitale de la France*). Sur le plan syntaxique, Moreau (1976) (citée par Kleiber, 1981) décrit des tests qui permettent de distinguer le sujet de l'attribut, notamment le clivage :

(59) a. C'est Paris qui est la capitale de la France.
b. * C'est la capitale de la France qui est Paris.

et le questionnement par *quel* :

(60) a. * Quelle est Paris ?
b. * Quelle est la capitale de la France ?

Sur le plan sémantique et référentiel, Kleiber analyse la différence entre les deux expressions en termes de différence de force identificatoire (nous avions déjà évoqué cette théorie de Strawson au chapitre 1). Ainsi, le terme auquel s'attache une « présupposition d'existence » est Paris : « on sait par là-même quel est le référent de *[Paris]*, ce qui n'est pas le cas pour l'expression *[la capitale de la France]*. Celle-ci au contraire peut avoir plusieurs occurrences. » Pour s'en convaincre, qu'on songe à une expression comme *La capitale de la France, quelle qu'elle soit (Paris, Lutèce, etc.)...* comparée à * *Paris, quel qu'il soit...*

La question est cependant plus compliquée pour les cas « d'énoncés *a est b* avec un prédicat relationnel d'égalité à deux arguments (*cf. Trieste est Vienne* dans le sens de "Trieste et Vienne désignent la même chose") » (Kleiber, 1981, p. 116). Kleiber écarte d'emblée l'étude de ces phrases « qui ne posent pas de problème d'interprétation ». Nous nous servirons donc des analyses de Frege (1971c [1892], p. 129) qui oppose le *est* copulatif (*i.e.* d'attribution) et le *est* qui « a le rôle du signe arthmétique d'égalité » et exprime une « égalité ». Dans le premier cas, *être* relie un « nom d'objet » (c'est-à-dire, dans nos termes, une expression référentielle) et un « terme conceptuel »[31] (qui est, d'un point de vue syntaxique, le prédicat) :

(61) [L'Étoile du matin]$_i$ est une planète.

Dans le second cas, *être* relie deux « noms d'objet » (*i.e.* deux expressions référentielles) :

(62) [L'Étoile du matin]$_i$ est [Vénus]$_i$.

On retrouve ici la relation à deux arguments que décrivait Kleiber. Pour Frege, le test qui permet de distinguer les deux cas est la convertibilité. Nous avons vu à l'instant qu'elle n'était pas suffisante (c'est-à-dire qu'il ne suffit pas de pouvoir dire *Vénus est l'Étoile du matin*), mais l'application des tests de Moreau (1976) cités par Kleiber (1981) montrent que, en effet, les deux expressions sont sur le même plan (ont la même force référentielle, dirait Kleiber) :

[31] On rappelle que la conférence citée s'intitule *Concept et Objet*.

(63) a. * C'est l'Étoile du matin qui est Vénus.
b. * C'est Vénus qui est l'Étoile du matin.
(64) a. * Quelle est Vénus ?
b. * Quelle est l'Étoile du Matin ?

ou avec des villes :

(65) a. * C'est Lutèce qui est Paris.
b. * C'est Paris qui est Lutèce.
(66) a. * Quelle est Paris ?
b. * Quelle est Lutèce ?

Nous pensons donc que, dans les cas de stricte égalité, il faudrait annoter « l'attribut » (qui n'est plus, alors, un attribut, puisque *est* n'est plus une copule, mais un prédicat à deux arguments et équivalent à *est égal à*) comme expression référentielle. Cependant, dans un souci d'homogénéité, nous ne l'avons pas fait ; de toute façon, le nombre d'occurrences concernées est insignifiant :

(67) [L'Étoile du matin]$_i$ est Vénus.

On peut cependant se demander ce qu'il faut faire des pronoms relatifs qui ont pour antécédent un SN attribut ; nous ne les avons pas annotés (c'est également le choix de Charolles, 2007) :

(68) [L'Étoile du matin]$_i$ est une planète qui est très petite.

Par contre, nous avons annoté les autres référents qui pouvaient apparaître *dans* un attribut, comme des compléments du nom :

(69) [L'Étoile du matin]$_i$ est une planète [du système solaire]$_j$

3 Bornage des maillons

Nous allons voir maintenant quelles sont les limites que nous avons données aux maillons.

3.1 Prépositions

Nous n'avons pas inclus l'éventuelle préposition dans le maillon :

(70) Le petit chat de [la voisine]$_i$

sauf dans les cas d'amalgames, où il nous est impossible[32] de séparer les deux :

(71) Le petit chat [du voisin]$_i$

[32] Il existe des corpus qui le font, mais le texte s'en retrouve alors modifié.

3.2 Relatives

Nous avons évoqué plus haut les différents types de relatives en nous interrogeant sur l'opportunité d'annoter les pronoms relatifs. Il est clair que les relatives déterminatives sont « intégrée[s] au SN » (Charolles, 2007) : leur intégration dans le maillon ne fait donc pas de doute. Quant aux autres types de relatives, nous les avons également intégré dans le maillon, de même que nous avons annoté tous les types de pronoms relatifs.

3.3 Appositions

La GMF (p. 343) définit les appositions de la façon suivante : « Mis en position détachée, le nom, le groupe nominal et l'adjectif ont une interprétation appositive. Ils sont au GN... ce que le complément circonstanciel est au reste de la phrase : un constituant périphrastique ». On pourrait donc ne pas les inclure dans le maillon, d'autant que, lorsque l'apposition est un nom, il y a coréférence avec le nom tête (« les deux GN sont dans un rapport d'identité référentielle », GMF, p. 354), comme dans :

(72) [Barack Obama, *l'actuel président des États-Unis*]$_i$, est en voyage diplomatique en Inde...

Nous avons cependant été obligé de les inclure pour des raisons techniques, puisque nous n'avons pas eu la possibilité de faire des annotations discontinues :

(73) [Barack Obama, *l'actuel président des États-Unis*, [qui]$_i$ est en voyage en Inde]$_i$, a prononcé un discours...

Ici, pour pouvoir inclure la relative dans le maillon, il nous faut aussi inclure l'apposition. (Notons que la relative est ici « appositive », mais nous parlons ici des appositions nominales, pour lesquelles il y a coréférence : *l'actuel président* est coréférent à *Barack Obama*, ce qui n'est pas le cas de la relative.) C'est un problème que nous retrouverons pour certaines prédications secondes et certains circonstants (voir la section « Compléments et circonstants », page 75).

Nous avons donc choisi d'inclure toutes les appositions nominales, afin de faciliter l'annotation et de rendre la classe plus homogène. Cependant, « les appositions nominales suivent toujours leur GN de rattachement, sauf celles qui sont dépourvues de déterminant et qui peuvent précéder le GN sujet » (GMF, p. 355), du type *Consul, Napoléon a fait...* Nous excluons ces dernières, *i.e.* nous ne les incluons pas dans le maillon, parce que nous les considérons davantage comme un circonstant, car comme l'explique la GMF, « il serait plus juste de dire que le GN apposé correspond à l'attribut d'une phrase à verbe *être* (mieux : d'une proposition attributive réduite, donc sans copule) dont le sujet serait le GN de rattachement ».

3.4 Groupes

Certaines expressions référentielles coordonnées donnent chacune lieu à une chaîne différente. C'est le cas par exemple dans :

[Pierre]$_i$ et [Marie]$_j$... [Il]$_i$... [Elle]$_j$

Mais nous avons parfois dû inclure plus d'éléments dans le maillon que nécessaire, pour des raisons techniques (pas d'annotations discontinues possibles[33]), comme dans l'exemple suivant, où nous trouvons d'abord la chaîne « des étudiants de deuxième [année] » puis celle « des étudiants... de quatrième année » :

(74) Peu de différences existent entre les zones centrales des représentations [[des étudiants de deuxième et de quatrième année...]$_i$]$_j$ (T4)

Nous aurions créé un troisième maillon incluant les deux groupes d'étudiants si cela avait donné lieu à une chaîne (ce qui n'était pas le cas) :

(75) Peu de différences existent entre les zones centrales des représentations [[[des étudiants de deuxième et de quatrième année...]$_i$]$_j$]$_k$ (T4)

Voici un autre exemple :

(76) [[Les capitaines hommes]$_i$ et femmes]$_j$ présentent des taux d'acceptations presque identiques... En revanche, il apparaît que [les capitaines hommes]$_i$ acceptent avec un peu plus de facilité que [les femmes]$_j$ des étudiants issus de formations différentes. (T2)

3.5 Rapide récapitulatif

En guise de bilan, nous présentons un rapide récapitulatif qui rappelle nos principaux choix. Malgré les symboles hérités des grammaires algébriques, il ne s'agit pas d'une telle grammaire. Les maillons annotés sont entourés de crochets ; les parenthèses représentent un regroupement ; un symbole peut être présent zéro ou une fois (dans ce cas, il est suivi par un point d'interrogation), zéro ou plusieurs fois (suivi par le signe $*$), une ou plusieurs fois (suivi par le signe $+$). Les coordinations ne sont pas prises en compte ici.

- verbe à l'actif (sauf impératif) :
  `[GN]? V ([GN]|PREP [GN])*`
- verbe à l'impératif :
  `[ø] V ([GN]|PREP [GN])*`
- verbe au passif :
  `[GN] AUX V (PREP [GN])*`
- nom non-prédicatif :
  `[N (PREP [GN])]`
- nom prédicatif et syncatégorématique :
  `[N (PREP [GN])*]`
- adjectif non prédicatif ou relative (quel que soit son type), toujours avec un nom :
  `[A* N (A|[PRO_RELATIF] RELATIVE)*]`
- adjectif prédicatif (toujours avec un nom dans ces études) :
  `[N [A (PREP [GN])+]+]`

[33] Elles seraient en fait possibles avec notre script d'annotation, ainsi qu'avec Analec, en ajoutant des propriétés qui identifieraient les différentes parties d'une même annotation. Mais la complexité qui en résulterait serait trop grande pour qu'une telle approche soit envisageable.

4 Propriétés linguistiques

Après avoir vu la façon dont nous avons choisi de borner les maillons, nous allons maintenant décrire les différentes propriétés linguistiques que nous avons annotées.

4.1 Catégorie grammaticale

Nous avons considéré quatre grands groupes de catégories grammaticales, parfois subdivisés en sous-classes.

Nous avons d'abord envisagé l'annotation des noms propres. Nous n'avons pas créé de sous-classes, mais nous avons au contraire regroupé tout ce qui peut être considéré comme un nom propre, quelle que soit la composition de l'expression ou sa détermination. Ainsi, nous considérons

(77) la Grotte des Treilles (T1)

comme un nom propre, et non comme un SN défini. En effet, les noms propres ont un comportement particulier : ce sont des SN (relativement) figés, dont aucun élément ne peut être commuté par un synonyme (78b), et qui n'admettent pas d'expansions supplémentaires ((78c) et (79b)) :

(78) a. la Grotte des Treilles
b. * la Caverne des Treilles
c. * la grande Grotte des Treilles

(79) a. le Petit Chaperon Rouge
b. * le beau (belle ?) Petit Chaperon Rouge

C'est pourquoi nous avons fait de ce groupe un groupe prioritaire (c'est-à-dire qu'un SN défini nom propre, comme « la Grotte des Treilles », sera classé comme un nom propre et non comme un SN défini).

Il existe cependant des cas où les noms propres peuvent « constituer le mot-tête d'un GN » (GMF, p. 316). Ils sont alors précédés d'un déterminant (s'ils n'en ont pas déjà un) et, dans la plupart des cas, une expansion supplémentaire (GMF, p. 317) :

(80) Le Barack Obama des années 1990...

(81) Le Petit Chaperon Rouge de Perrault (par opposition au Petit Chaperon Rouge des frères Grimm)...

Nous n'avons pas rencontré de noms propres utilisés de cette façon dans notre corpus. Si nous en avions rencontré, nous les aurions alors classés plutôt dans les SN définis.

Le deuxième grand groupe est celui des SN à tête nominale (par opposition aux SN à tête pronominale), que nous avons subdivisé selon la détermination. Outre les traditionnels SN défini, indéfini, possessif et démonstratif, nous avons créé une classe pour les SN sans déterminant, par exemple *le concept d'[interdisplinarité]*$_i$.

Le troisième groupe est celui des SN à tête pronominale. Nous l'avons subdivisé selon le type de pronom. D'abord la classe des pronoms personnels, dans laquelle nous incluons *y* et *en*,

qui peuvent commuter avec des pronoms personnels (*je les mange/j'en mange/j'y mange*), ce qui tend à montrer qu'ils peuvent appartenir à la même classe. Nous avions hésité à créer une classe à part pour ces deux pronoms, mais y avons renoncé afin de ne pas multiplier les classes. De plus, il est facile de séparer ces pronoms *a posteriori* et de façon automatique, puisqu'il suffit de chercher les maillons qui ne se composent que de *en* ou de *y*. Nous avons aussi créé une classe pour les pronoms réfléchis (lorsque l'objet est coréférent avec le sujet, comme dans *nous nous demandons*) et réciproques (lorsqu'il y a coréférence *et* distribution, comme dans *nous nous battons*). Cette fois, nous avons été obligé de créer une classe à part pour garder la compatibilité avec Democrat, qui ignore ces pronoms (une classe séparée permet de facilement supprimer ces pronoms). Nous avons aussi créé une classe pour les pronoms zéro, comme nous l'avons expliqué ci-dessus (65). Enfin, nous avons annoté les traditionnels pronoms relatifs et démonstratifs.

Pour finir, nous avons considéré le déterminant possessif comme une classe à part, puisque ce n'est pas un SN, et qu'il ne pouvait entrer dans aucune des catégories précédentes. Il est particulier en ce qu'il a une référence tout en n'étant pas un SN[34].

Pour récapituler :

- nom propre,
- SN à tête nominale :
  - sans déterminant,
  - défini,
  - indéfini,
  - possessif,
  - démonstratif ;
- SN à tête pronominale :
  - pronom personnel,
  - pronom réfléchi,
  - pronom relatif,
  - pronom démonstratif,
  - pronom zéro ;
- déterminant possessif.

Ce schéma est un compromis : il a des limites. La première est l'existence d'un chevauchement entre les classes : un SN défini, par exemple, pourra être dans la classe des noms propres ou celle des SN définis. Il n'y a pas d'ambiguité puisque celle des noms propres est prioritaire, mais cela révèle l'intrusion d'un critère à caractère sémantique (nom propre) dans une distribution qui est sinon syntaxique.

Il existe par ailleurs une autre limite : les SN à tête nominale sont classés selon des critères à caractère référentiel (le déterminant indique l'actualisation et donc la référence), alors que les SN à tête pronominale sont classés selon le type du pronom, donc selon des critères formels. Ainsi, si on compare :

[34] Le déterminant possessif n'est pas le seul dans ce cas : les déictiques peuvent aussi être référentiels sans être des SN, comme *aujourd'hui* ou *ici*. Nous n'en avons cependant pas rencontré dans les articles de recherche.

(82) a. Je mange des haricots.
b. J'en mange.
(83) a. Je mange les haricots.
b. Je les mange.

on constate que (82a) est indéfini (partitif) alors que (83a) est défini, mais leur équivalent pronominalisé (82b) et (83b) entrent tous deux dans la catégorie des pronoms personnels. Et même si nous avions séparé les pronoms *y* et *en* dans une classe à part (dans ce cas, (82b) n'aurait pas été dans la même classe que (83b)), cette classe aurait été définie de façon on ne peut plus formelle, ce qui aurait encore ajouté au mélange des critères.

4.2 Fonction grammaticale

4.2.1 Modèle théorique

Plutôt que de suivre la grammaire scolaire, qui multiplie les catégories, nous avons préféré adopter un modèle théorique fondé sur la notion d'arbre syntaxique, sans toutefois entrer dans le détail de la grammaire générative. Nous avons ainsi isolé le sujet et les circonstants, puis rattaché les compléments qui restaient à l'une des classes « complément du verbe », « de l'adjectif » ou « du nom » en fonction de leur dépendance. Nous avons aussi traité les cas spéciaux des titres et parenthèses.

4.2.2 Compléments et circonstants

La façon dont nous avons annoté les compléments demande quelques explications. Nous présenterons séparément les compléments de verbes d'une part et le compléments de noms et d'adjectifs d'autre part.

Pour les verbes

En ce qui concerne les verbes, la principale difficulté réside dans la distinction entre compléments et circonstants. Il y a plusieurs moyens de définir ce qu'est un complément : restriction syntaxique (*i.e.* complément d'objet), restriction sémantique (*i.e.* structure argumentale), différence entre modifieurs et compléments à la façon des générativistes (Delaveau, 2001, pp. 58–61, Carnie, 2013, pp. 175–195)...

Nous avons considéré ici que seuls les compléments « sélectionnés » (Choi-Jonin et Delhay, 1998, p. 215) ou « sous-catégorisés » (GMF, p. 230, Delaveau, 2001, p. 45) étaient des compléments du verbe. La GMF (p. 260) et Delaveau (2001, p. 62) proposent des tests pour faire la différence entre compléments et circonstants ; nous nous sommes cependant appuyé sur les tests que Choi-Jonin et Delhay (1998, pp. 215 *sq.*) proposent[35], notamment le test d'omission et celui consistant à insérer « et cela » entre le verbe et le syntagme à tester. Dans ce dernier test, « cela » est une anaphore résomptive qui reprend, et donc permet d'isoler, le prédicat et ses compléments sélectionnés. Par exemple, dans (84a), « une lettre » est un complément sélectionné du verbe parce qu'il ne peut pas en être séparé par « et cela » (84b) :

[35] Ces tests sont repris de Tellier, 1995.

(84) a. Je t'écrirai une lettre.
b. * Je t'écrirai et cela une lettre.

Mais « chez ma belle-mère » n'est pas un complément sélectionné dans (85a), car on peut séparer ce syntagme du verbe par « et cela » (85b) :

(85) a. J'ai déjà mangé chez ma belle-mère.
b. J'ai déjà mangé et cela chez ma belle-mère.

Pour les noms et les adjectifs

La question de la distinction entre compléments sous-catégorisés et circonstants se pose aussi pour les noms et les adjectifs, notamment ceux « à caractère verbal », *i.e.* les noms et adjectifs prédicatifs, et les participes utilisés comme adjectifs. Nous avons cependant été obligé de les traiter autrement que les verbes, et d'inclure les circonstants dans le maillon.

D'abord, parce que ces circonstants ne dépendent ni du verbe ni de la phrase, mais bien du nom ou de l'adjectif annoté :

(86) [Les feuilles tombées *en automne*]$_i$ se sont envolées au printemps
(87) Nous avons ici utilisé [l'échelle de compétences politiques validée lors de l'étude 1]$_i$ (T0)

Dans (86), *en automne* se rattache indubitablement à *tombées*, tandis que le circonstant de phrase est *au printemps* : les deux circonstants ne sont pas au même niveau hiérarchique.

Parfois, ces indications sont des « modifieurs » au sens de Delaveau (2001, p. 60), c'est-à-dire des éléments certes facultatifs, mais qui dont la présence « restreint la désignation du constituant qu'ils modifient », et donc participent à la référence. En effet, (88a) et (88b) ne sont pas équivalents :

(88) a. [Le changement d'heure en automne] est (dés)agréable.
b. [Le changement d'heure] est (dés)agréable.

Ensuite, parce qu'il est techniquement impossible d'exclure ces circonstants de niveau hiérarchique secondaire. Imaginons en effet qu'on veuille annoter l'exemple suivant :

(89) a. Les feuilles que j'aimais tant voir ont été ramassées.

Puisque nous considérons, en suivant la GMF, que la relative est une expansion, nous l'incluons dans le maillon :

(90) a. [Les feuilles [que]$_i$ j'aimais tant voir]$_i$ ont été ramassées.

Maintenant, imaginons qu'on veuille annoter cette phrase :

(89) b. Les feuilles *tombées en automne* et que j'aimais tant voir ont été ramassées.

Ici, *en automne* est un circonstant de la prédication seconde :

(91) Les feuilles (sont) tombées, et cela en automne...

Si l'on veut inclure la relative dans le maillon comme nous l'avons fait dans (90a), on est obligé d'inclure aussi la prédication seconde, et donc le circonstant :

(90) b. [Les feuilles *tombées en automne* [que]$_i$ j'aimais tant voir]$_i$ ont été ramassées.

Ni Analec, ni Glozz, ni le script que nous avons écrit ne permettent, en effet, des annotations discontinues. (92) présente un exemple authentique de ce phénomène :

(92) La variabilité isotopique observée ici plaide en faveur de [pratiques alimentaires *variées au sein de ce groupe* pour lesquelles il reste encore difficile de cerner toutes les ressources alimentaires qui ont pu contribuer]$_i$. (T1)

4.2.3 Autres fonctions

Nous avons créé une classe « autre » pour les cas qui ne rentraient pas dans les classes précédentes, comme le complément du comparatif. Pour la GMF (p. 233), ce complément entre dans un système corrélatif rattaché aux compléments circonstanciels. Néanmoins, comme il gravite autour d'un adjectif, on pourrait aussi le considérer comme un complément de l'adjectif. Nous n'avons pas pris parti, et l'avons mis dans la classe « autre ».

Nous avons enfin créé une dernière classe, regroupant les titres et les parenthèses qui n'entrent pas dans une phrase. Par contre, lorsque la fonction est clairement déterminée (par exemple un complément du nom), alors nous avons considérée cette fonction. Par exemple, dans le titre :

(93) [Cadre théorique de [la recherche]$_i$]$_j$

nous avons mis *Cadre...* dans la classe « autre » et *la recherche* dans celle des compléments du nom.

4.2.4 Le problème des déterminants possessifs

Comme nous l'avons déjà évoqué dans la section « Principes et compromis », page 63, nous avons considéré les déterminants possessifs comme des compléments du nom, puisque *son chat* est l'équivalent de *le chat de lui*, sans toutefois créer une classe à part en ce qui concerne la fonction, classe qui aurait été propre au déterminant possessif. C'est justement parce qu'elle aurait été propre au déterminant que nous ne l'avons pas créée : elle aurait été inutile (puisqu'on peut récupérer l'information simplement par le codage de la catégorie grammaticale (voir la section « Catégorie grammaticale », page 73)) et aurait ajouter une classe supplémentaire, peu peuplée et donc parasitaire pour les études statistiques.

4.2.5 Le problème des participiales

Dans les participiales du type :

(94) Une fois [l'équipe]$_i$ créée, le capitaine peut accueillir des coéquipiers... (T2)

(95) Une fois [l'invitation]$_i$ lancée, l'étudiant ainsi invité peut accepter ou refuser [l'invitation]$_i$ (T2)

nous avons considéré que *l'équipe* et *l'invitation* étaient des *sujets* des participes.

4.3 Expansions

Les expansions sont des « éléments facultatifs » (GMF, p. 242) qui gravitent autour des éléments essentiels. La GMF distingue entre « ajouts » (p. 242), expansions du verbe, et « modifieurs » (p. 342), expansions du nom. Nous garderons cependant le terme général d'expansion, puisque nous n'avons ici à faire qu'à un seul type (expansions du nom).

La GMF (p. 342) fait une distinction entre les modifieurs « nécessaires à l'identification du référent » (rapport déterminatif) et les modifieurs « qui ne restreignent pas l'extension du nom » (rapport explicatif). Nous ne faisons pas cette différence, dans un souci d'homogénéité syntaxique (la distinction est sémantique) et pour ne pas multiplier les classes, selon le principe évoqué plus haut.

La GMF (ch. 7) propose la liste d'expansions du nom suivante :

- adjectif épithète,
- nom épithète, apposé directement à un autre nom, sans préposition ni déterminant (« une tarte maison »),
- groupe prépositionel :
  - complément du nom (N1 de N2),
  - détermination qualitative antéposée (« cet imbécile de Pierre »),
  - déterminant composé (« un tas de problèmes »)[36] ;
- propositions :
  - subordonnée relative,
  - subordonnée complétive (« l'idée que... »),
  - construction infinitive introduite par *à* (« un livre à lire »).

De tous ces éléments, nous avons formé des classes à partir de la nature grammaticale de la tête du groupe qui fait l'expansion :

- adjectif, y compris les participes,
- nom :
  - soit en complément du nom, *i.e.* séparé du nom expansé par une préposition,
  - soit « épithète » comme dans *la plateforme Studyka* (T2) ou *le projet Canal Seine-Nord Europe* (T4) ;
- proposition ; nous ne trouvons ici, à quelques exceptions près (*une demande pour rejoindre une équipe* (T2)), que des relatives.

Il faut noter que nous avons suivi la GMF (p. 348) et considéré que dans des expressions comme *le métier de vétérinaire rural* (T4), *le thème général « du travail »* ou *le concept d'interdisciplinarité* (T2), *de vétérinaire rural*, *du travail* et *d'interdisciplinarité* étaient des compléments du nom (qui indiquent une « catégorisation discursive », au sens où *vétérinaire rural* vient catégoriser *métier*) et non des appositions (ce qui est pourtant l'analyse traditionnelle, avec l'exemple de *la ville de Paris* (GMF, p. 348)).

Nous avons dû inclure les noms apposés pour les raisons techniques mentionnées dans la section « Compléments et circonstants », page 75 :

[36] Puisqu'il s'agit d'un déterminant, on peut douter qu'il s'agisse aussi d'une expansion...

(96) [la Grotte I des Treilles, site éponyme de ce complexe archéologique]$_i$ (T1)

(97) [une liste de 25 mots ou expressions, les mêmes que ceux utilisés dans le questionnaire précédent]$_i$ (T4)

Cependant, il nous a semblé important de ne pas les mettre sur le même plan que les autres expansions. En effet, sur le plan syntaxique elles sont *détachées*, ce qui les éloigne des autres expansions, et, sur le plan sémantique, elles sont coréférentielles au nom auquel elles se rattachent : ce sont donc davantage des « proposition[s] attributive[s] réduite[s] » (GMF, p. 354) que des expansions à proprement parler. Nous les avons donc annotées de façon différente, même si, ce faisant, nous mélangeons deux critères (catégorie grammaticale et position dans le syntagme).

D'un point de vue pratique, nous avons attribué à chaque type d'expansion un code (*a* pour l'adjectif, *n* pour le nom, *s* pour les propositions (*s* comme *subordonnées*) et *p* pour les appositions). Si un nom a plusieurs expansions, alors nous indiquons plusieurs plusieurs codes (éventuellement les mêmes, comme dans les coordinations, par exemple *le petit et gros chat* aura le code « aa »). Cela permet de compter le nombre d'expansion.

Il faut indiquer pour finir que nous ne prenons en compte que les expansions de même niveau, ce qu'un générativiste appellerait une « domination directe » (Carnie, 2013, pp. 120 *sq.*). Ainsi, dans :

(98) [le petit chat de [ma vieille voisine]$_i$]$_j$

nous comptons comme expansion de *chat* « petit » et « de ma vieille voisine », mais non « vieille ».

4.4 Propriétés linguistiques non annotées

Nous n'avons annoté ni le genre ni le nombre car nous avons estimé que ce n'était pas pertinent pour les référents abstraits et les groupes.

Il y a deux façons d'annoter le genre : soit le genre grammatical du mot (ce qui peut généralement se retrouver automatiquement), soit l'identité sexuelle du référent (ce qui implique l'usage d'un genre « neutre » pour les inanimés). Le genre des mots est, en français, arbitraire : nous avons donc estimé que leur étude n'était pas utile pour notre travail. Il pourrait être intéressant d'étudier le genre grammatical par opposition au sexe, mais la plupart de nos référents sont soit abstraits (donc sans sexe), soit des groupes (donc mêlant hommes et femmes).

Quant au nombre, il est très généralement possible de le retrouver automatiquement (par analyse de la tête lexicale, que nous avons annotée : voir ci-dessous). Nous n'en avons cependant pas fait l'étude, car nous avons pensé qu'il est relativement prévisible en fonction du type de référent (un référent abstrait sera singulier, un groupe souvent pluriel, etc.).

5 Propriétés non linguistiques

Nous avons également annoté, pour chaque maillon, le mot tête du syntagme. Cela est très rapide et ne demande que peu d'effort avec notre script d'annotation (chapitre 3), et a trois avantages. D'abord, ce marquage permet le calcul de la distance intermaillionnaire dans les cas de maillons imbriqués :

(99) [Le petit chat [que]$_i$ j'ai adopté quand [il]$_i$ était tout jeune]$_i$

Il n'est pas possible de calculer la distance entre les maillons dans ces cas-là, et c'est pourquoi nous avons calculé la distance intermaillonnaire en comptant le nombre de tokens *entre deux têtes de syntagmes*. Dans l'exemple, il s'agit du nombre de tokens entre *chat*, *que* et *il*.

Par ailleurs, le marquage de la tête permet de calculer automatiquement le coefficient de stabilité lexicale, dont nous donnons une description plus complète page 82. Sans le marquage de la tête, ce calcul aurait été impossible à faire avec plus de trois milliers de maillons.

Enfin, comme nous venons de le constater dans le titre précédent, la tête permet de calculer automatiquement le nombre, et, dans une large mesure, le genre.

5.1 Rapide récapitulatif

Nous rappelons ici l'ensemble du schéma d'annotation (avec les codes cités dans l'annexe A) que nous avons utilisé pour ce travail.

Catégorie :

- nom propre (n),
- SN à tête nominale : sans déterminant (t), défini (d), indéfini (i), possessif (p), démonstratif (m) ;
- SN à tête pronominale : pronom personnel (s), pronom réfléchi (f), pronom relatif (r), pronom démonstratif (o), pronom zéro (z) ;
- déterminant possessif (e).

Fonction : sujet (s), complément du verbe (v), complément de l'adjectif (a), complément du nom (n), circonstant (c), autre (t), titre ou parenthèse (h).

Tête syntaxique (index et forme).

Expansion (zéro, une ou plusieurs fois) : adjectif (a), nom (n), proposition (s), apposition (p).

Chapitre 5

Une première étude exploratoire : annotation d'une sélection de référents saillants

1 Introduction

Malgré les difficultés que nous avons évoquées au chapitre 1 pour identifier des chaînes de référents abstraits, nous proposons ici les résultats d'une première étude exploratoire, au cours de laquelle nous avons annoter cinq articles scientifiques.

1.1 Le choix des référents

Nous avons décidé d'annoter les référents sur tout le texte, sans tenir compte du découpage interne au texte (parties, paragraphes : ce que nous ferons dans la deuxième étude, au chapitre 6) ; c'est pourquoi nous avons sélectionnés et annotés les référents les plus saillants au niveau du texte.

Le choix des référents résulte d'une procédure systématique. Nous avons inclus systématique comme référents l'« auteur », la « recherche » et l'« article ». Nous donnons ci-dessous une définition de ces termes.

Nous avons ensuite sélectionné des référents supposés saillants ou importants à partir des mots-clés, du titre et du résumé.

Nous avons également lemmatisé et étiqueté (au niveau de la catégorie grammaticale) chaque texte avec l'outil TreeTagger : cela nous a permis d'observer les fréquences les plus élevées des lemmes nominaux (sans la lemmatisation, nous aurions été dépendants des variations morphologiques). Cependant, il s'est avéré que le choix de référents en fonction des fréquences

est difficile, puisque, d'une part, un même référent peut être désigné par différents lemmes, et que, d'autre part, deux expressions contenant le même lemme peuvent référer à des entités différentes. Nous avons ainsi cherché à créer une chaîne pour un capitaine d'équipe (texte T2), alors que les différentes expressions « capitaine d'équipe » dans le texte ne renvoient en fait pas au *même* référent.

Enfin, nous avons utilisé l'outil AntConc pour le calcul de n-grammes, qui nous ont souvent permis de découvrir d'autres référents.

1.2 Méthode de calcul de quelques indicateurs

Outre les propriétés annotées pour chaque maillon, nous avons calculé automatiquement des indicateurs qui caractérisent une chaîne dans son ensemble, notamment sa longueur (*i.e.* le nombre de maillons qui la composent), la distance intermaillonnaire et les coefficients de stabilité lexicale et formelle. Nous développons dans cette section le mode de calcul des trois derniers.

1.2.1 La distance intermaillonnaire

La distance intermaillonnaire est la distance moyenne entre les maillons. Nous n'avons pas pu prendre en compte la distance entre les bornes antérieures et postérieures des maillons, puisque cela nous aurait obligé à ignorer les maillons imbriqués (par exemple les pronoms relatifs) :

(100) [Le petit <u>chat</u> [<u>que</u>]$_i$ j'ai adopté]$_i$

Nous avons donc choisi de calculer la distance entre les têtes syntaxiques des maillons (soulignées dans l'exemple (100)). Il faut noter qu'il serait aussi possible, en l'absence de balisage de la tête syntaxique, d'utiliser la borne antérieure du maillon (c'est-à-dire le premier mot du maillon) ; il s'agit là en effet d'une bonne approximation, surtout pour des chaînes qui ont souvent une distance intermaillonnaire de plusieurs dizaines ou centaines de mots. En effet, nous avons constaté que pour 32 % des maillons, le premier élément *est* la tête syntaxique (c'est par exemple le cas d'un pronom personnel ou relatif), et que pour 66 %, la tête est le deuxième élément (précédé d'un article le plus souvent, comme dans *le <u>chat</u>*). La distance maximale entre la tête et la borne antérieure est de cinq mots, et cela n'arrive que pour un seul maillon (sur les près de 2000 annotés). La marge d'erreur est donc négligeable.

1.2.2 Les coefficients normalisés de stabilité

Le coefficient de stabilité a été défini par Perret (Perret, 2000, p. 17) comme la division, « pour un référent donné (un personnage), [du] nombre total d'anaphores nominales par le nombre de désignations différentes ». Il s'agit d'un indicateur important pour les chaînes de référence. Néanmoins ce mode de calcul ne permet pas de calculer des moyennes, ni même de comparer des chaînes entre elles.

Stabilité lexicale

Nous avons donc créé un indicateur normalisé, c'est-à-dire dont les valeurs sont comprises entre 0 (ou 0 % : aucune stabilité, toutes les expressions sont différentes) et 1 (ou 100 % : stabilité maximale, toutes les expressions sont identiques). Il se définit par la formule :

$$1 - \frac{x-1}{n-1}$$

pour $n > 1$ et $1 \leq x \leq n$, n et x étant des entiers, où x est le nombre de termes *différents* (le « nombre de désignations différentes » de Perret) et n est le nombre *total* de termes (le « nombre total d'anaphores nominales »).

Pour le coefficient normalisé de stabilité *lexicale* (CNSL), les termes dont il est question sont des noms : les pronoms sont donc ignorés. Dans ce cas, il peut arriver que n soit égal à 1 (par exemple dans « Le chat... il... il... »). Le coefficient est alors défini à 1.

Prenons un exemple. Soit la chaîne suivante :

(101) Le petit chat... Il... Il... Ce félin... Il... Le chat... Il... Le chat... Il...

La chaîne (101) contient neuf maillons, mais seulement quatre noms ; on a donc $n = 4$. Trois de ces noms sont identiques, nous avons donc deux noms différents (« chat » et « félin ») ; on a donc $x = 2$. Le coefficient de stabilité lexicale est donc :

$$1 - \frac{2-1}{4-1} = 1 - \frac{1}{3} \approx 0.667$$

La chaîne suivante, par contre, a un coefficient de 0, puisque toutes les désignations de Pierre sont différentes :

(102) Pierre... Le jardinier... Il... M. Dupont... qui...

Au contraire, dans la chaîne (103), le coefficient est de 1, puisque toutes les désignations (c'est-à-dire toutes les têtes syntaxiques de maillon) sont les mêmes :

(103) Le jardinier... Il... Le jardinier bien élevé... qui... Le grand jardinier... que...

Stabilité formelle

Pour le coefficent normalisé de stabilité *formelle* (CNSF), les termes dont il est question sont les expressions telles qu'elles apparaissent dans le texte, ce qui permet de prendre en compte la détermination et les modifieurs. Autrement dit, si deux expressions ne diffèrent que par l'ajout d'un modifieur (« le chat » *vs* « le petit chat »), le coefficient de stabilité *lexicale* sera le même, mais pas le coefficient de stabilité *formelle*.

Lors du calcul de ce dernier indicateur, nous avons ignoré systématiquement les pronoms, sauf si la chaîne n'était constituée *que* de pronoms (ce qui arrive notamment dans les chaînes dont les maillons sont des pronoms de première ou deuxième personne). Sans cela, le résultat

aurait été 0 (car $x = 0$), et il aurait été impossible de faire la différence entre une chaîne sans nom et une chaîne où tous les noms diffèrent.

Reprenons la chaîne (101) ci-dessus. Nous avons vu qu'elle contenait quatre noms ; donc $n = 4$. Nous regardons désormais le syntagme nominal dans son ensemble, et pas seulement sa tête syntaxique : nous avons trois formes différentes (« Le petit chat », « le félin » et « le chat ») ; on a donc $x = 3$. Le coefficient de stabilité formelle est donc :

$$1 - \frac{3-1}{4-1} = 1 - \frac{2}{3} \approx 0.334$$

On note donc que la stabilité formelle est dans ce cas plus faible, ce qui s'explique par l'ajout de l'adjectif « petit » au premier maillon. De manière générale, on peut s'attendre à ce que la stabilité formelle soit toujours plus faible que la stabilité lexicale.

Nous avons calculé ces indicateurs automatiquement, ce qui nous oblige à faire quelques remarques supplémentaires. Pour la stabilité lexicale, nous avons utilisé la tête syntaxique, que nous n'avons pas lemmatisée : une différence en genre et/ou en nombre peut donc entraîner une différence lexicale non voulue. Cela est cependant très rare dans les cas des chaînes de référence. Ensuite, toujours en ce qui concerne la stabilité lexicale, nous avons utilisé la stabilité formelle si la chaîne n'était constituée que de pronoms (ce qui évite le cas aberrant où $x = 0$). Enfin, en ce qui concerne cette fois la stabilité formelle, nous avons ignoré les articles définis et les amalgames au début des expressions, afin d'éviter de fausser le calcul avec des formes contractées (« du chat » *vs* « le chat »).

1.3 Autres indicateurs utilisés

Outre la distance intermaillonnaire et les coefficients normalisés de stabilité lexicale et formelle, nous avons calculé :

- la longueur moyenne des chaînes (en maillons). Il s'agit simplement du nombre moyen de maillons dans une chaîne,
- la longueur moyenne des maillons (en tokens). Il s'agit du nombre moyen de tokens par maillon,
- la longueur moyenne des paragraphes[37] (en tokens). Il s'agit du nombre moyen de tokens par paragraphe,
- la densité globale (en maillons par paragraphe). Il s'agit du rapport entre le nombre de maillons (toutes chaînes confondues) et le nombre de paragraphes. Ce rapport indique le nombre de maillons qu'on peut s'attendre à trouver dans un paragraphe. Par exemple, si la densité est de 2 maillons par paragraphe, alors on trouvera (en moyenne) deux maillons (toutes chaînes confondues) dans un paragraphe,
- le nombre moyen de paragraphes couverts (en paragraphes). Il s'agit du nombre moyen de paragraphes dans laquelle une chaîne apparaît. Par exemple, si le nombre de paragraphes couverts par une chaîne est de 10, alors la chaîne apparaîtra dans 10 paragraphes,
- la densité moyenne des paragraphes couverts (en maillons par paragraphe). La densité des paragraphes couverts par une chaîne est le rapport entre le nombre de maillons de

[37] Utilisé seulement au chapitre 6.

cette chaîne et le nombre de paragraphes couverts par cette chaîne. Par exemple, si le rapport est égal à 3, alors dans chaque paragraphe où la chaîne considérée apparaît, on pourra espérer y trouver trois maillons (en moyenne).

2 Dénombrements et statistiques

Puisque les chaînes annotées ont été pré-sélectionnées (c'est-à-dire qu'il n'y a pas eu d'annotation systématique de toutes les chaînes, comme ce sera le cas au chapitre 6), les statistiques globales (tableau 1) ne sont pas représentatives. Nous n'en ferons donc qu'une analyse très limitée.

Le nombre des référents sélectionnés est compris entre 15 et 22, selon les textes.

La longueur moyenne des chaînes (*i.e.* le nombre de leurs maillons) oscillent entre 12/14 pour les textes T1 et T3, et 24/28 pour les trois autres. Cette répartition correspond à la différence entre sciences naturelles et sciences humaines, mais nous verrons que, dans le détail, il y a beaucoup de variation entre chaque texte ; aussi, il est difficile d'opposer les disciplines à ce stade.

Le texte T3 semble être un peu à la marge : sa densité (*i.e.* le rapport entre le nombre de maillons et le nombre de paragraphes) est très faible (3.38), par comparaison avec les autres textes (au moins 6.49), mais cela résulte peut être du fait que nous ayons annoté peu de chaînes (15).

Ce qui semble caractéristique des textes scientifiques, et ce qu'on verra dans les pages qui suivent, c'est le coefficient de stabilité lexicale : il est très haut, c'est-à-dire qu'il y a très peu de changement dans la désignation des référents.

Notons aussi qu'en moyenne, il y a 140 tokens par paragraphe. Cette indication est utile notamment pour interpréter la distance intermaillonnaire de certaines chaînes.

3 Études par type de chaîne

Nous avons réparti les différentes chaînes en classes, selon le type de référent et les difficultés d'annotations rencontrées. Nous distinguerons :

- la chaîne de l'auteur,
- les chaînes « recherche » et « article »,
- les entités nommées et les référents définis,
- les ensembles,
- les noms massifs,
- les références génériques,
- les noms abstraits,
- les noms prédicatifs,
- les « ensembles flous ».

	nombre de tokens	nombre de chaînes	longueur moyenne des chaînes (en maillons)	distance intermaillonnaire	coefficient normalisé de stabilité formelle	coefficient normalisé de stabilité lexicale	nombre de maillons	longueur moyenne des maillon (en tokens)	densité globale (en maillons par par.)	nombre total de par.	nombre moyen de par. couverts	densité moyenne (par. couverts) (en maillons par par.)
T0	6541	15	28.93	273.7	0.65	0.91	434	3.37	7.89	55	13.67	1.98
T1	6209	20	14.6	500.2	0.38	0.77	292	4.76	8.34	35	6.65	2.01
T2	7255	22	28.27	239.17	0.53	0.88	622	3.13	8.08	77	12.96	2.16
T3	4410	15	12.4	330.54	0.49	0.84	186	3.3	3.38	55	8.87	1.36
T4	8731	17	24.82	345.79	0.46	0.79	422	2.7	6.49	65	9.82	2.97
corpus	33146	89	21.98	339.41	0.5	0.84	1956	3.49	6.82	287	10.39	2.09
introduction	7690	88	5.63	115.99	0.57	0.77	495	3.64	8.25	60	2.63	2.15
méthodologie	5579	65	4.91	100.13	0.52	0.79	319	3.42	4.62	69	2.17	2.5
résultats	11341	75	6.89	250.83	0.57	0.84	517	3.29	5.02	103	4.01	1.65
discussion	6611	69	6.88	156.08	0.57	0.84	475	3.6	13.19	36	2.51	2.83
conclusion	1925	47	3.19	58.24	0.61	0.82	150	4.23	7.89	19	1.64	1.99

Tab. 1 : Statistiques globales de la première étude exploratoire. Voir les sections « Méthode de calcul de quelques indicateurs », page 82, et « Autres indicateurs utilisés », page 84, pour le calcul des différents indicateurs.

3.1 La chaîne de l'auteur

L'étude de la chaîne de l'auteur est intéressante parce qu'elle s'inscrit dans de nombreuses recherches sur l'expression de l'auteur scientifique (dans une perspective sociologique : Pontille, dans une perspective linguistique : Grossmann, 2010, dans la perspective des textes scientifiques : Loffler-Laurian, 1980 ou Fløttum, 2006b).

C'est donc une chaîne que nous avons annoté systématiquement. On pourrait penser que son référent est relativement stable et présent tout au long du texte. En fait, elle présente plusieurs difficultés.

D'abord, les maillons qui la composent sont presque exclusivement le pronom personnel de première personne « nous » et le déterminant possessif correspondant « notre, nos ». Mais il est parfois difficile de savoir à qui réfère « nous ». Si dans la phrase :

(104) Nous avons choisi d'utiliser cette mesure du burnout pour deux raisons majeures. (T0)

il n'y a pas de doute possible, il est plus difficile de savoir ce qu'englobe « nous » dans :

(105) À la lecture du tableau 1, nous remarquons que sept challenges... (T2)

S'agit-il des auteurs, ou bien d'un « nous » qui incluerait aussi le lecteur ? La même question se pose pour « on ». Parfois, il ne désigne clairement pas l'auteur, comme dans :

(106) Ces filières peuvent être combinées deux à deux : on parle alors de filière « mixte ». (T4)

Ailleurs, il désigne un étudiant (et donc pas l'auteur) :

(107) lorsque l'on fait partie d'une équipe interdisciplinaire... (T2)

Müller-Gjesdal, 2013 dresse un tableau des différentes valeurs que peut prendre « on » (tableau repris de Fløttum, Jonasson et Norén, 2007) : sur les six valeurs identifiées, deux n'incluent pas l'auteur : l'une qui renvoit au lecteur, l'autre qui désigne des personnes qui ne sont ni l'auteur ni le lecteur (comme les exemples (106) et (107) ci-dessus).

Il faut aussi se demander, comme le font Landragin et Tanguy (2014), si « les alternances de *on* inclusif et de *nous* sont... coréférentielles au sens strict ». Les deux auteurs parlent en termes de « référence floue » (qui peut être inclusive, exclusive, générique ou totalement floue), et distinguent « plusieurs degrés de coréférence » : la « coréférence stricte » lorsque les référents sont identiques, la « coréférence inclusive » lorsque l'un inclut l'autre, et la « coréférence floue » lorsque « les deux référents sont des groupes flous et l'intersection entre les deux est possible, tout en restant potentiellement floue elle-même » (p. 111).

Pour faciliter l'annotation, nous avons avons choisi d'inclure systématiquement le pronom de première personne dans la chaîne, même dans les cas douteux comme en (105). Nous avons ainsi inclus tous les pronoms « on », sans faire de distinction de degrés de référence ou de coréférence. Par contre, nous n'avons inclus le pronom « on » que lorsqu'il était clair qu'il désignait ou incluait l'auteur, et non lorsqu'il référait au lecteur ou à une autre personne.

Si c'est le pronom (« nous » ou « on ») qui est utilisé dans la majorité des cas, lorsque l'auteur (ou les auteurs) devient le propos du texte, il est nominalisé en « auteur ». C'est le cas dans le texte T2 :

(108) Deux des trois auteurs du présent article sont co-fondateurs de Studyka... Pour cette raison, l'accès aux données de la plateforme fût libre et illimité. (T2)

Mais le paragraphe continue par :

(109) Les auteurs ont eux-mêmes procédé aux extractions... (T2)

sans que l'on sache lesquels des (un, deux ou trois) auteurs ont procédés aux extractions des données.

On constate de très grandes différences entre les textes, comme le montre le tableau 2. Seuls deux textes ont véritablement une chaîne « auteur » (T0 et T2). Il faut de plus noter que dans le texte T4, il y a plus de pronoms « on » que de pronoms « nous ».

Chaîne	L CR	DI	L MA	DENS	CNSF	CNSL	PAR
(T0) Auteur	79	80.18	1	2.55	0.96	0.96	31
(T1) Auteur	9	577.88	1	1.5	0.75	0.75	6
(T2) Auteur	74	89.12	1.07	2.24	0	1	33
(T3) Auteur	6	706	1	1.2	0.6	0.6	5
(T4) Auteur	7	314.5	1	1.75	0.83	0.83	4
Moyenne	35	353.54	1.01	1.85	0.63	0.83	15.8

Tab. 2 : Statistiques pour la chaîne « auteur ». Voir la section « Abréviations et symboles », page v pour la signification des abréviations.

Il ne semble pas y avoir de différence entre les disciplines, puisque, des deux textes qui ont une chaîne auteur conséquente, l'un appartient aux sciences humaines, l'autre aux sciences naturelles.

Par ailleurs, on remarque que la longueur des maillons est très réduite (de l'ordre d'un token) : cela s'explique par la composition de la chaîne « auteur », qui est constituée uniquement (ou presque) par des pronoms.

On peut aussi opposer l'usage des pronoms et celui des déterminants possessifs, les deux seules catégories vraiment représentées. Le pronom est de loin le plus utilisé (75% des cas), presqu'exclusivement en fonction sujet, rarement en tant que complément du verbe.

Les maillons de cette chaîne se répartissent différemment selon les parties (tableau 3).

partie	nb de maillons
introduction	63
méthodologie	10
résultats	49
discussion	30
conclusion	23

Tab. 3: Répartitions des maillons de la chaîne « auteur ».

D'autres études ont analysées les marques de l'auteur dans les articles de format IMRaD, notamment Heslot (1983) et Régent (1980), citées et résumées par Regent (1992), qui prennent en compte les pronoms « nous » et « on », mais aussi, comme nous, les déterminants possessifs ; et Müller-Gjesdal (2013), qui s'est seulement intéressé aux pronoms « nous » et « on » dans un corpus de cinquante textes médicaux. Les résultats diffèrent des nôtres. Alors que Heslot et Régent trouvent une forte présence de l'auteur dans l'introduction et la discussion, « qui constituent la trame argumentative de l'article destinée à convaincre et informer le public plus large » (Regent, 1992), Müller-Gjesdal, trouve la plupart des pronoms dans la partie « méthode et résultats » (elle n'a pas séparé les deux), et, dans une moindre mesure, dans la discussion. Quant à nous, nous en trouvons la plupart dans l'introduction et les résultats... Cela est peut-être dû au fait que seuls deux de nos textes ont une chaîne « auteur » de quelque ampleur. Ou peut-être que cela est révélateur de pratiques différentes, selon les auteurs, les disciplines, ou même les époques.

En conclusion, cette chaîne, si elle est présente, apparaît tout au long du texte, mais n'est pas répartie également sur toutes les parties. Nous pouvons faire l'hypothèse qu'elle sera presque toujours composée de pronoms sujets et, dans une moindre mesure, de déterminants possessifs.

3.2 Les chaînes « recherche » et « article »

Nous avons essayé de séparer, d'une part, les expressions qui réfèrent à la recherche qui précède l'écriture de l'article, ce qui inclut, par exemple, les « hypothèses », les « problématiques » et « les réflexions », et, d'autre part, les expressions qui réfèrent à l'article lui-même, exposé de la recherche précédente. Cette distinction peut cependant sembler un peu arbitraire, c'est pourquoi nous traitons ici des deux chaînes en même temps.

Nous avons eu quelques difficultés à définir ce qu'on pouvait faire entrer dans la « recherche ». Ainsi, nous avons rejeté les expérimentations puisque si elle font partie de la recherche, elles ne renvoient pas, d'après nous, au même référent. Cependant, nous avons inclus les cas d'études (avec un sens plus abstrait que celui d'« expérimentation ») successivement présentées dans le texte T3. Nous nous heurtons là à la difficulté des référents abstraits, dont les limites sont floues. Puisque nous n'avons pas pu établir de règles pour accepter ou rejeter un terme, nous avons fait appel à notre intuition, en nous fondant surtout sur la synonymie lexicale.

Chaîne	L CR	DI	L MA	DENS	CNSF	CNSL	PAR
(T1) Recherche	4	853	2.25	1.33	0.33	0.67	3
(T2) Article	27	253.73	2.78	1.5	0.84	1	18
(T2) Recherche	24	289.26	3.25	1.6	0.29	0.48	15
(T3) Recherche	5	595.5	4.8	1	0.75	1	5
(T4) Recherche	12	723.91	2.5	1.71	0.55	0.82	7
Moyenne	14.4	543.08	3.11	1.42	0.55	0.79	9.6

Tab. 4: Statistiques pour les chaînes « recherche » et « article ».

Seuls les textes T2, et T4 contiennent un nombre significatif de maillons pour ces deux chaînes (tableau 4). La chaîne « article » n'apparaît que pour le texte T2, qui est aussi le texte pour lequel la chaîne « recherche » est la plus grande. C'est aussi l'un des deux textes pour lesquels la chaîne « auteur » était importante. Il est à ce titre étonnant que les chaînes « recherche » et « article » n'apparaissent pas dans le texte T0 : c'est en partie dû à notre choix d'exclure les expérimentation de ces référents.

Dans le texte T2, les maillons de la chaîne « article » se répartissent bien entre les parties, à l'exception de la partie « résultats ». Souvent, les expressions donnent des indications de repérage : « l'article est organisé... », « dans la suite de l'article... ».

La chaîne « recherche », par contre, apparaît principalement dans l'introduction et dans la conclusion, d'abord pour décrire l'élaboration de la recherche (« observations », « analyse », « problématique », « question de recherche », « étude »), ensuite pour en rappeler les limites (« *ses* principales limites »).

Au contraire, dans le texte T3, la chaîne « recherche » apparaît dans la partie « résultats », qui expose ce à quoi la recherche a aboutit, dans la partie « discussion ». De façon significative, le mot récurrent est ici « étude », alors que le vocabulaire était bien plus diversifié dans le texte T2, comme le montre le coefficient de stabilité lexicale (1 contre 0.48).

Dans le texte T4, la chaîne est développée surtout dans la partie « discussion ». Ici encore le terme employé est « étude » (un coefficient de stabilité de 0,82). Cela pourrait suggérer une corrélation entre les parties et les diversité lexicale : le vocabulaire est bien moins diversifié lorsqu'il est question de la recherche dans les parties « résultats » et « discussion » (T3 et T4) que lorsqu'il en est question dans les autres parties (T2), même si nous n'avons pas assez de données pour savoir si cette hypothèse est valide.

Dans l'ensemble des textes, les maillons de ces chaînes sont surtout des SN définis et démonstratifs, et, dans le texte T2 où la chaîne de l'auteur est présente, possessif (« notre étude »).

Les fonctions se partagent entre sujet et complément de noms comme *objectif, résultat, contexte, cadre théorique, limite*.

En conclusion, il faudrait dire que le repérage de ces chaînes n'est pas facile et laisse une grande part à l'intuition de l'annotateur. Cependant, elles semblent avoir des comportement différents non seulement en fonction des textes, mais aussi en fonction des parties.

3.3 Les entités nommées et les référents définis

Nous abordons maintenant un autre type de chaînes : les entitées nommées et les référents définis. Nous mettons dans cette classe les référents qui existent en dehors du discours et sont (ou du moins peuvent être) connus du lecteur en dehors de l'article. Par exemple, la Grotte I des Treilles (texte T1) est un référent qui existe indépendamment de l'article. Nous excluons de cette catégorie les référents abstraits (page 101) et « massifs » (page 99), qui sont rassemblés dans des catégories à part.

L'annotation de ces chaînes ne présentent pas de difficultés majeures, puisque le référent est unique, stable et généralement concret. Deux occurrences seulement ont posé problème :

dans le texte T1, il est difficile de savoir si « des Treilles » dans les « hommes des Treilles » réfèrent à la grotte ou à la civilisation. Cette ambiguïté se rapporte à celles des ensembles qu'on verra plus loin : s'agit-il des sujets étudiés (première interprétation) ou à l'ensemble de la population du Groupe des Treilles (seconde interprétation) ?

S'ils sont relativement faciles à annoter, ces référents sont cependant très rares dans les articles scientifiques. On pourrait les répartir sur une échelle qui mesurerait le degré d'ancrage et de stabilité dans la conscience des auteurs et des lecteurs. Par exemple, la Grotte I des Treilles et le Groupe des Treilles (T1) sont des entités clairement définies, bien ancrées et très stables dans la conscience des archéologues, puisque la première est un site géographique, et le deuxième une culture du néolithique français, qui se développe autour des Grottes des Treilles.

La plateforme Studyka est un site internet utilisé par les auteurs du texte T2 pour récupérer des données sur le comportement d'étudiants. Néanmoins, cette entité est moins bien définie : qu'est-ce qu'un site Internet ? Ce peut être un ensemble de pages, des outils et/ou des services, une interface, des données, mais aussi un réseau, ou (dans le cas qui nous préoccupe) une entreprise, etc.

On peut aussi inclure dans cette catégorie la « filière rurale » du cursus universitaire pour devenir vétérinaire (texte T4). Elle semble définie hors de l'article, et pourtant n'a pas le caractère plus ou moins concret des trois référents précédents. L'annotation de ce référent a posé problème lorsqu'il a été question de « filière mixte », mélange de deux filières. Nous avons exclu ces filières mixtes, même si cela est un peu arbitraire.

L'échelle de Ferris est une échelle de mesure du burnout utilisée dans le monde anglophone (texte T0). Il est intéressant de la comparer avec sa version française, puisque c'est justement cette version française qu'établit l'article T0. La version française n'existe donc pas antérieurement à l'article, mais elle est censée être utilisée par la suite.

Nous avons également annoté le « projet du Canal Seine-Nord Europe », qui est un projet de fouilles archéologiques. Nous n'avons pas annoté les mentions du canal lui-même, mais celle du projet, ce qui est en fait un référent certes défini, mais à un niveau moindre que les autres (et que ne le serait le canal lui-même).

Chaîne	L CR	DI	L MA	DENS	CNSF	CNSL	PAR
(T0) Échelle de Ferris	31	192.9	4.45	3.1	0.54	0.92	10
(T0) Échelle Française	23	267.23	4.52	2.56	0.19	0.75	9
(T1) Grotte des Treilles	31	196.5	4.58	2.07	0.79	0.9	15
(T1) Groupe des Treilles	12	471.27	3.83	1.5	0.91	0.91	8
(T2) Plateforme Studyka	47	132.83	2.4	1.81	0.81	0.93	26
(T3) Projet du Canal	8	556.43	7	1	0.71	0.71	8
(T4) Filière Rurale	24	351.7	2.92	1.85	0.41	0.73	13
Moyenne	22.88	316.06	4.11	1.91	0.59	0.81	11.75

Tab. 5: Statistiques pour les entités nommées et les référents définis.

La longueur des chaînes et leur distance intermaillonnaire (tableau 5) sont très variables, et ne semblent pas être des critères discriminants. Il faut d'ailleurs noter que la chaîne du ca-

nal, avec ses huit maillons, n'est pas très représentative. Par contre le coefficient de stabilité lexicale paraît être un indicateur relativement fiable : plus le référent est stable, moins ses expressions varient. Les mentions de la Grotte et du Groupe des Treilles ne varient que très peu (0,9), alors celles de référents comme de la version française de l'échelle de Ferris ou la filière vétérinaire rurale ont une plus grande variation (0,75 et 0,73), surtout si on considère la variation formelle (0,19 et 0,41), ce qui suggère d'importantes différences dans la détermination.

Mais c'est surtout par la catégorie lexicale qu'on peut opposer ces référents. Plus un référent est défini en-dehors de l'article, et plus ses expressions seront de type « nom propre » ou « SN défini ». Le taux est de plus de 90 % pour la Grotte, le Groupe, le projet du Canal, et presque autant pour la plateforme Studyka. Pour les autres référents, par contre, les catégories grammaticales sont plus diverses. Ainsi, pour l'échelle de Ferris, s'il y 65 % de SN définis, il y a aussi beaucoup de SN indéfinis, démonstratifs ou sans déterminant, avec quelques pronoms. La version français de cette échelle, est, quant à elle, surtout mentionnée à l'aide de SN possessifs (surtout avec le possessif de première personne : « notre adaptation de l'échelle », « notre version », etc.), alors que le taux de SN définis tombe à 13 %.

Il y a une forte dichotomie en ce qui concerne l'expansion. Pour prendre les extrêmes : alors qu'aucune des mentions du Groupe des Treilles (que nous avons considérés comme une expression figée, un nom propre, non expansé) ne sont expansées, environ une mention de la version française de l'échelle de Ferris sur deux l'est.

Les fonctions principales de ces référents sont celles de sujet, de complément du nom et de circonstant, mais il ne semble pas y avoir de différence discriminante.

Ces chaînes ont-elles un comportement différent selon les parties dans lesquelles elles apparaissent ? Pour l'ensemble, on peut établir le tableau 6.

partie	nb de maillons
introduction	49
méthodologie	39
résultats	31
discussion	43
conclusion	14

Tab. 6: Répartitions des maillons des chaînes d'entités nommées et de référents définis.

On remarque donc une plus forte présence de ces chaînes dans l'introduction, ce qui s'explique aisément : cette partie est le lieu où sont présentés les différents éléments ; il semble assez peu utile de les répéter par la suite, puisque ce sont des référents stables, bien ancrés dans la conscience des lecteurs. Et pourtant ils se maintiennent assez bien sur les parties suivantes, sauf dans la conclusion, ce qui est un peu étonnant, puisqu'on aurait pu attendre une reprise des éléments de l'introduction.

Dans le détail, la Grotte se trouve dans toutes les parties, entre quatre et huit occurrences à chaque fois. Le Groupe des Treilles, par contre, a la plupart de ses occurrences dans l'introduction. Cela s'explique par les différentes fonctions de ces référents : alors que le Groupe sert à situer la recherche dans un domaine archéologique spécifique (ce qui est fait en intro-

duction), la Grotte, elle, sert à qualifier les sujets (des restes humains) de l'expérimentation. Comme l'article décrit, sur toute sa longueur, cette expérimentation, il est normal que ce référent ce retrouve dans toutes ses parties. Les maillons de ces chaînes sont dans la plupart des cas (deux occurrences sur trois) compléments du nom : ils qualifient les sujets (pour la Grotte) et la population et les sites (pour le Groupe) étudiés. Une occurrence sur cinq est complément circonstanciel.

Dans le texte T2, la plateforme Studyka est surtout employée dans l'introduction (15 occurrences sur 47) et la méthodologie (19). Au contraire les autres parties sont très pauvres (résultats : 4, discussion : 6 et conclusion : 2). Dans la méthodologie, un tiers des occurrences sont sujets (« Studyka a organisé 11 challenges ») et un autre tiers circonstants (« les étudiants inscrits sur cette même plateforme »).

Le projet archéologique du Canal Seine-Nord Europe trouve des occurences dans chaque partie, mais leur nombre est toujours très faible, ce qui en limite l'intérêt.

Enfin, la filière rurale des études vétérinaires est présente non seulement dans l'introduction (7 occurrences sur 24) et la méthodologie (5 occurrences), mais aussi, et surtout, lors de la discussion (12 occurrences). Dans cette partie, la moitié des occurrences sont compléments du nom, tendance qui ne se dégage ni en introduction, ni en méthodologie.

En conclusion, on peut retenir que si ces référents sont faciles à repérer et à annoter, ils sont rares dans les articles scientifiques. On peut les répartir sur une échelle qui mesure leur ancrage et leur stabilité dans la conscience des lecteurs, et leur position sur cette échelle se reflète par différents paramètres linguistiques, comme la stabilité de leurs mentions (au niveau du lexique et de l'expansion), ou encore leur catégorie grammaticale. Certains traits semblent se dégager en fonction des parties (ils se concentrent plus facilement dans certaines parties), et c'est une piste qu'il faudrait explorer en annotant davantage de textes.

3.4 Les ensembles

Les articles de format IMRaD utilisent beaucoup les expérimentations statistiques. C'est le cas de tous les textes que nous avons annotés. Or les études statistiques cherchent à prouver l'existence (ou l'absence) de certaines propriétés chez un groupe, à partir d'un échantillon, donc d'un groupe plus restreint. C'est dire que les ensembles sont très fréquents dans les articles de ce format.

Les expressions qui réfèrent à des ensembles sont parmi les plus difficiles à annoter. Cette difficulté résulte surtout de la multiplication des sous-ensembles.

Une partie des problèmes est résolue en prenant une définition intensionnelle des ensembles, et non une définition extensionnelle. Par exemple, une population d'étudiants (textes T2 et T4) changera entre le début et la fin de l'expérimentation (il y a des abandons, des nouveaux venus). Les échantillons eux-mêmes ne sont pas stables : dans le texte T2, certaines équipes qui participent à des challenges, et qui font donc partie de l'échantillon, abandonnent à mi-parcours. Les chercheurs gèrent différemment les cas d'abandons en fonction de leur méthodologie.

Mais il y a plus ennuyeux : le texte T2 prend comme échantillon plusieurs équipes s'affrontent dans une sorte de concours. La plupart des équipes sont donc éliminées au cours de l'expérimentation. L'ensemble des équipes qui participent à un challenge n'est donc pas le même, extensionnellement, au début à la fin de l'expérimentation, puisqu'il ne contient pas les mêmes éléments. Ce problème est contourné en prenant une définition intensionnelle, c'est-à-dire en ne considérant que la définition (les équipes qui participent à un challenge), et non la liste des membres.

Il y a ensuite des problèmes de vocabulaire : dans quelle mesure « l'échantillon » ou « les participants » (texte T4) sont-ils des mentions du même référent ? Puisque du point de vue de l'analyse statistique (or nos textes sont des études statistiques), ils sont très proches (les participants forment l'échantillon), nous les avons mis dans la même chaîne.[38]

La différence entre ipséité et identité est bien connue des philosophes. Si j'ai *la même voiture* que mon voisin, on suppose (par inférence pragmatique) qu'il s'agit du même modèle (identité), pas du même exemplaire. Si j'habite *le même appartement* que mes parents, on suppose qu'il s'agit au contraire du même exemplaire (ipséité). L'inférence ne suffit plus, par contre, si j'affirme que j'ai *la même voiture* que mes parents (même exemplaire ou même modèle ?).

On trouve un problème similaire dans le texte T4 :

(112) Les mêmes termes que ceux utilisés dans le questionnaire précédent. (T4)

Du point de vue philosophique, il y a identité, et non ipséité. Cela rejoint la distinction entre phrase et énoncé : « on ne prononce jamais deux fois le même énoncé ». Mais alors, les mots de chaque questionnaire formeraient des ensembles différents, et les participants répondraient chaque fois à des questions différentes... ce qui est contre-intuitif. Nous avons donc considéré qu'il s'agissait du même référent, et avons inclus « les mêmes termes » dans la chaîne des termes du « questionnaire précédent ».

Cet exemple, de même que le problème de la définition intensionnelle des ensembles, montre qu'il y a une différence entre ce que nous considérons comme un référent et ce qu'un philosophe considérerait comme un référent. Dans tous les cas, nous avons veillé à suivre l'intention de l'auteur et notre intuition de lecteur plutôt que la logique philosophique. En effet, comme le rappelle cp :Asher-1993, nous ne passons pas notre vie cognitive à résoudre des problèmes de référence. Intension/extension et identité/ipséité sont des problèmes qui demandent une réflexion que le lecteur cible (qui est un sociologue ou un archéologue, et non un linguiste ou un philosophe) ne fait pas.

C'est dans cette optique, également, que nous avons considéré qu'« étudiants » dans « le choix[39] des étudiants » (texte T4) référaient à l'ensemble des étudiants et non pas à un sous-

[38] De même avons nous considéré, que *réponses* dans la phrase (110) :

(110) Là aussi, l'anonymat des réponses a été garanti. (T4)

était dans la même chaîne que *participants* dans la phrase (111) :

(111) L'anonymat des participants était garanti. (T4)

En effet, les deux phrases, qui se situent chacune à la fin de deux paragraphes consécutifs, se répondent. Mais il s'agit là d'une exception, et n'avons pas, sinon, considéré que les *personnes* et les *questions* renvoyaient au même référent.

[39] Choix de la filière de spécialisation : les étudiants vétérinaires peuvent choisir entre une filière "animaux domestiques", une filière "rurale" (*i.e.* animaux de production, comme les vaches), etc.

groupe. *Strico sensu*, en effet, tous les étudiants ne choisissent pas leur filière vétérinaire (par exemple « animaux domestiques » ou « animaux de production ») : seuls ceux de quatrième année le font, et encore seulement à la fin de l'année. On pourrait donc opposer les étudiants qui vont choisir (dans un futur plus ou moins proche), ceux qui s'apprêtent à choisir, ceux qui sont en train de choisir, ceux qui ont choisi, ceux qui reviennent sur leur choix, etc. Nous avons, ici encore, choisi ce qui est le plus intuitif, et considéré que « le choix des étudiants » concerne tous les étudiants faisant des études pour devenir vétérinaire.

Par contre, lorsqu'il était explicitement mentionné qu'ils ne s'agissaient que d'étudiants d'une année particulière (par exemple : « les étudiants de troisième année de cursus »), nous n'avons pas inclus la mention dans la chaîne de tous les étudiants.

Nous avons suivi un principe similaire lorsqu'il a fallu intégrer les « équipes non finalisées » dans les équipes qui participent à un challenge, parce qu'elles le peuvent au début (mais seulement au début : après, elles doivent être finalisée si elles veulent continuer, sinon elles sont éliminées). Il en va de même pour des expressions telles que :

(113) Le plus souvent, [les défis lancés par les entreprises]$_i$ doivent donner lieu à la création d'une entité nouvelle. (T2)

Manifestement, ce ne sont pas *tous* les défis qui sont concernés par cette règle, mais seulement la plupart, ce que souligne « le plus souvent ». Pour simplifier, nous avons, là encore, inclus l'expression dans la chaînes des défis.

Un problème similaire apparaît avec les référents génériques, dont il sera question encore ci-dessous. Peut-on considérer que les expressions suivantes renvoyent au même référent, qui serait un groupe ?

(114) a. Tous les oiseaux ont des ailes.
b. Les oiseaux ont des ailes.
c. Un oiseau a des ailes.
d. L'oiseau a des ailes.

L'exemple (113) montre que (114a) n'est pas équivalent à (114b). Par ailleurs, Corblin (1987) refuse de considérer que (114c) décrit un ensemble équivalent au quantificateur universel, ce qui signifie que (114a) et (114c) n'ont pas les mêmes conditions de vérité. Dans ce cas, on ne pourrait pas mettre « (tous) les oiseaux » et « un oiseau » dans la même chaîne, puisque les ensembles n'ont pas la même définition, même au niveau intensionnel. Nous avons décidé, certes un peu arbitrairement puisque même intuitivement la situation n'est pas si claire, de considérer que ces expressions renvoyaient toutes au même référent. Ce faisant, nous avons rencontrer le problème des variables liées (décrit plus loin), dont il est parfois difficile de faire la différence avec la référence générique.

Le problème de la phrase (114d) se rencontre par exemple dans :

(115) Studyka a organisé 11 challenges en ligne ayant pour chacun une ou plusieurs entreprises à l'initiative du challenge. (T2)

Ici encore, en restant cohérent avec notre choix précédent concernant l'indéfini, nous avons réunis « 11 challenges » et « du [= de le] challenge » dans la même chaîne.

Un autre problème concerne les expressions, très nombreuses dans les textes de notre corpus, du type « un quart des étudiants », « la majorité des sujets », « le reste de la population », « 3 % des hommes ». On peut considérer que ces expressions ne réfèrent, chacune, qu'à un seul ensemble. Et pourtant, pour considérer un quart de cent étudiants, soit 25 étudiants, on doit d'abord considérer l'ensemble des cent étudiants. En appliquant ce principe, nous aurions deux référents dans chacune de ces expressions (en décomposant l'amalgame pour être plus clair) :

(116) a. $[\text{un quart de } [\text{les étudiants}]_j]_i$
b. $[\text{la majorité de } [\text{les sujets}]_j]_i$
c. $[\text{le reste de } [\text{la population}]_j]_i$
d. $[\text{3 \% de } [\text{les hommes}]_j]_i$

Nous avons préféré la deuxième solution, celle qui compte deux référents dans chacune de ces expressions, parce que sinon les étudiants, par exemple, seraient cités sur de longs passages (par exemple lors de la description de tableaux), sans jamais faire partie d'aucune chaîne. En effet, un auteur peut décrire des données en disant, par exemple « 3 % des étudiants sont dans telle situation alors que 3,8 % sont dans telle autre situation » : ces deux mentions n'entreraient dans une aucune chaîne (à moins de retrouver la mention des « 3 % d'étudiants »), ce qui serait, à notre avis, contre intuitif.

Enfin, il est parfois difficile de faire la différence entre échantillon et population, notamment dans les parties « discussion » et « conclusion ». C'est le moment où l'auteur extrapole les résultats trouvés à partir de l'échantillon à toute la population. Mais alors que la différence est souvent très claire dans la partie « méthodogie », à aucun moment, dans les cinq textes annotés, l'auteur indique clairement cette extrapolation dans les résultats, la discussion ou la conclusion. On ne peut la comprendre que parce que l'on connaît par avance le fonctionnement d'une étude statistique. Ainsi, on ne sait pas, à la fin de l'article, si « les étudiants » réfèrent à l'échantillon sondé ou bien à la population générale des étudiants qu'on cherche à décrire. Il s'agit, selon les règles des études statistiques, des deux à la fois, puisqu'il y a extrapolation des résultats trouvés sur l'échantillon à la population générale. Nous avons choisi de considérer ces expressions comme la population générale, puisque c'est probablement là l'intention de l'auteur. Mais aucun élément linguistique n'indique cela, seule la connaissance du fonctionnement de telles études permet de faire ce choix.

Ce fait est bien souligné dans le texte T2, lorsque, discutant des limites de leur article, les auteurs comparent les deux ensembles d'étudiants :

(117) La première limite de l'article concerne la population d'élèves considérée dans l'article. Elle n'est pas représentative de l'ensemble des étudiants du supérieur. (T2)

Ici, des marques linguistiques (« *considérée* dans l'article », « *l'ensemble* des étudiants du supérieur ») permettent de bien faire la différence entre les deux populations.

Il existe d'autres cas qui relèvent du même problème, et qui semblent indécidables. Par exemple, dans le texte T1, on trouve :

(118) Une analyse comparative avec d'autres sites a été tentée dans le but d'identifier d'éventuelles particularités alimentaires [des sujets inhumés dans la Grotte I des Treilles]$_i$. (T1)

S'agit-il des seuls sujet examinés, sachant que tous les sujets inhumés dans la Grotte I n'ont pas servi à l'étude ? La question se complique lorsque l'on sait que la Grotte I est une grotte sépulcrale, qui sert de cimetière.

Tous les articles que nous avons annotés sont des études statistiques, mais tous ne font pas appels à des échantillons d'objets concrets : il peut s'agir de mesures. Ainsi l'étude T3 se fonde sur les valeurs de susceptibilité magnétique. Nous n'avons pas pu annoter ces ensembles de valeurs, car ils combinent les difficultés des ensembles avec celles des référents abstraits (décrits page 101). Nous nous sommes donc contenté des échantillons concrets, qu'il s'agisse d'humains, ou de restes humains.

Dans le texte T1, nous avons pris la population de la Grotte I des Treilles (la population, au sens statistique du terme), ainsi que les sujets humains et animaux de l'étude (les échantillons). Dans le texte T2, nous avons pris les étudiants en général (la population) et les étudiants de l'étude (l'échantillon). Nous avons aussi pris comme ensemble les équipes : toutes les équipes, et les sous-groupes des équipes inter-disciplinaires et mono-disciplinaires. Enfin, nous avons suivi l'ensemble des onze challenges. Dans le texte T4, nous avons suivi les étudiants en études vétérinaires (la population) et les participants aux expérimentations : tous les participants mais aussi différents sous-groupes (puisqu'il y a différentes expérimentations). Nous avons aussi inclus des groupes de questions (25 mots que les étudiants devaient jugés et 10 mots qui évoquaient leur représentation de la pratique de vétérinaire rural). Le tableau 7 résume ces choix.

Aucun regroupement ne semble pouvoir être fait avec ces données. La longueur et la distance intermaillonnaire sont très variables. Il y a des chaînes locales, comme les « 10 mots » que devaient trouver les étudiants du texte T4 : elle est brève et la distance intermaillonnaire est faible ; c'est parce qu'elle n'est évoquée que pendant un bref instant. On pourrait faire les mêmes observations pour la chaîne des « 25 termes », ou bien pour les sous-groupes de participants du texte T4. À l'autre extrême, la population du groupe des Treilles est tout aussi brève (six maillons), alors que la distance intermaillonnaire est près de cent fois plus grande. C'est parce qu'on la retrouve surtout en introduction (deux occurrences) et en conclusion (deux), avec deux mentions au milieu du texte. Cela pourrait être une piste pour opposer parmi les chaînes brèves celles qui sont locales (distance intermaillonnaire très brève) et celles qui sont globales (distance intermaillonnaire très longue), bien que les « sujets animaux » du texte T1 soit un cas un peu intermédiaire.

Le texte T2 présente une vue plus unifiée de ce type de chaînes. On décèle une certaine corrélation inverse entre la longueur de la chaîne et la distance intermaillonnaire : plus une chaîne a de maillons, plus ils sont rapprochés. Équipes et étudiants présentent des chaînes avec un nombre important de maillons, et la différence entre les équipes inter-disciplinaires (31 maillons) et les équipes mono-disciplinaires (7 maillons) trahit le thème principal du texte (l'interdisciplinarité). Néanmoins, il est difficile de tirer des conclusions sur la base d'un seul texte.

Chaîne	L CR	DI	L MA	DENS	CNSF	CNSL	PAR
(T1) Population Étudiée	6	1091.4	7.67	1.5	0	0.6	4
(T1) Sujets Animaux	6	749.4	3	3	0	0.4	2
(T1) Sujets Humains	21	224.95	4.62	2.63	0.39	0.72	8
(T2) Equipes Interdisciplinaires	31	86.17	2.77	2.58	0.81	0.96	12
(T2) Equipes Monodisciplinaires	7	329	3.57	1.17	0.6	0.8	6
(T2) Étudiants	57	123.66	1.75	4.38	0.84	0.95	13
(T2) Étudiants Studyka	75	84.49	2.37	2.68	0.69	0.9	28
(T2) Tous Les Challenges	25	245.33	4.2	1.47	0.26	0.74	17
(T2) Toutes Les Equipes	13	295.92	2.54	1.86	0.5	0.92	7
(T3) Bâtiments Néolithiques	8	394.71	3	1.14	0.43	1	7
(T3) Grandes Fosses	6	62	2.83	1.5	0	0.75	4
(T4) 10 Mots	7	9.33	3.71	7	0	0.5	1
(T4) 203 Participants	9	246.75	2	4.5	0.6	0.8	2
(T4) 25 Mots	14	448.39	4.64	3.5	0	0.71	4
(T4) 330 Participants	10	151.78	3.4	3.33	0.5	0.83	3
(T4) 38 Participants	7	183.17	1.71	3.5	0.33	0.67	2
(T4) Étudiants	58	143.28	2.88	2.64	0.61	0.84	22
(T4) Participants	27	213.19	2.11	2.45	0.65	0.87	11
Moyenne	21.5	282.38	3.27	2.82	0.4	0.78	8.5

Tab. 7: Statistiques pour les ensembles.

En ce qui concerne la catégorie grammaticale, aucune tendance claire ne se dégage. Presque toutes ces chaînes ont plus de 50 % de SN définis, avec quelques indéfinis et quelques pronoms personnels. Les expansions sont très variables. La fonction qui prédomine est celle de complément du nom, ce qui n'est pas étonnant (« 5 % des participants », etc.).

La répartition des chaînes entre les différentes parties est conforme à notre attente : l'introduction comporte surtout les maillons des chaînes qui ont pour référent la population (au sens statistiques), alors que la méthodogie et les résultats contiennent surtout les chaînes des participants (échantillons). La discussion et la conclusion contiennent les deux, en tenant compte de la difficulté d'annotation (échantillon ou population) signalée plus haut. Nous pouvons illustrer ce schéma avec le texte T2 (tableau 8).

partie	population	échantillon
introduction	18	9
méthodologie	1	13
résultats	0	28
discussion	29	13
conclusion	7	13

Tab. 8 : Répartition des chaînes de la population et de l'échantillon statistique dans le texte T2 (en nombre de maillons).

En conclusion, on peut dire que les chaînes qui réfèrent à des ensembles posent de vraies difficultés d'annotation. Elles s'opposent surtout par leur distribution dans les différentes parties

IMRaD, mais ne semblent pas, sinon, pouvoir se regrouper par des caractéristiques linguistiques. Cela est peut-être dû au manque de données, ou à de mauvais choix d'annotations (peut-être faudrait-il être plus inclusif, ou l'être au contraire moins).

3.5 Les noms massifs

Nous avons annoté certains référents massifs. Ils sont relativement faciles à annoter, puisqu'ils changent très peu. La seule difficulté est celle de savoir, par exemple, si « azote atmosphérique » a le même référent que « azote ». Cela revient à se demander si l'eau de Strasbourg et l'eau de Paris sont la même eau. Nous avons choisi de suivre ici la visée communicative de l'auteur : si l'intention de celui-ci est d'opposer l'eau de Strasbourg à celle de Paris, alors il conviendrait de ne pas attribuer à ces deux expressions le même référent. Dans le cas de l'azote, l'opposition entre « azote atmosphérique » et celui qui vient de la terre n'est pas significative (du moins pas à l'échelle du texte), et donc nous n'avons pas séparé les référents de ces entités. Il en va de même pour le collagène : « le collagène des herbivores », « le collagène des prédateurs », dans ces expressions nous voyons essentiellement du collagène. Il faut cependant garder à l'esprit que les expansions permettent de faire des distinctions référentielles (voir pour un exemple la section « Les noms prédicatifs », page 104), qu'on se trouve, ici, devant le problème fondamental de l'individusation et de l'identification des référents abstraits, et qu'aucun des choix d'annotation que l'on peut faire (mettre dans la même chaîne ou non les « deux » collagènes) ne semble satisfaisant.

Le tableau 9 montre les référents annotés.

Chaîne	L CR	DI	L MA	DENS	CNSF	CNSL	PAR
(T1) Azote	20	276.63	1.5	2.22	0.84	0.9	9
(T1) Carbone	10	461	1.4	2	0.78	0.89	5
(T1) Collagene	7	454.67	4.29	1.17	0.17	0.83	6
(T3) Matière Organique	10	356.22	2.1	1.43	0.89	0.89	7
Moyenne	11.75	387.13	2.32	1.7	0.67	0.88	6.75

Tab. 9: Statistiques pour les noms massifs.

Ces chaînes semblent assez homogènes, tant du point de vue de la distance intermaillonnaire, que de leur longueur (par comparaison avec les autres chaînes, qui peuvent avoir près de 90 maillons), et de leur stabilité formelle et lexicale. La stabilité formelle relativement faible du collagène vient de ses expansions (« des herbivores, des prédateurs, des consommateurs », etc.). Il faut noter que l'azote et le carbone sont souvent représentés par leur symbole chimique ; ainsi l'azote est tantôt désigné par « azote », tantôt par « 15N » et tantôt par « N » seulement.

Les maillons de ces chaînes sont surtout des SN définis (« du collagène ») ou sans déterminants (« les teneurs isotopiques en azote »). Le taux d'expansion est très stable (soit le terme est expansé sur tout le texte, soit il ne l'est pas du tout), sauf pour le collagène, pour les raisons déjà évoquées. La fonction la plus fréquente, et de loin, est celle de complément du nom (entre 80 et 90 % pour l'azote, le carbone et la matière organique), à l'exception du collagène (57 %) qui

est aussi dans la catégorie « autre complément », qui regroupe notamment les compléments des comparatifs et les parenthèses.

Il y a une véritable opposition selon les parties : on les trouve presqu'exclusivement dans la méthodologie et les résultats (tableau 10).

partie	nb de maillons
introduction	4
méthodologie	16
résultats	20
discussion	5
conclusion	2

TAB. 10: Répartitions des maillons des chaînes de noms massifs.

Ces référents sont donc faciles à annoter, présents sur tout le texte, mais avec des fréquences beaucoup plus importantes dans les parties « méthodologie » et « résultats ». Ils ont un comportement homogène du point de vue linguistique : catégorie grammaticale, fonction, expansion, et varient très peu au cours du texte. Cependant, nos données ne sont pas suffisamment nombreuses pour pouvoir tirer de réelles conclusions.

3.6 Les références génériques

Nous avons annoté deux référents « génériques », sans savoir si l'on peut véritablement parler de « chaîne de référence générique ». Il s'agit de la « surface décapée » du texte T3, et du « vétérinaire rural » du texte T4 (tableau 11).

Chaîne	L CR	DI	L MA	DENS	CNSF	CNSL	PAR
(T3) Surface Décapée	12	388.09	3.25	1.2	0.36	0.82	10
(T4) Vétérinaire Rural	20	434.47	1.9	2.86	0.4	0.8	7
Moyenne	16	411.28	2.58	2.03	0.38	0.81	8.5

TAB. 11: Statistiques pour les référents génériques.

Les expressions qui réfèrent à ces référents peuvent être au singulier :

(119) a. Les fonctions majeures assurées par [le vétérinaire rural]$_i$... (T4)
b. La carte de champ total correspond à la carte d'anomalies mesurées à 30 cm au-dessus de [la surface décapée]$_i$. (T3)

ou au pluriel :

(120) a. [Les vétérinaires ruraux]$_i$ assurant un rôle de lien social de proximité auprès de population. (T4)
b. La mesure de paramètres magnétiques sur [les surfaces décapées]$_i$... (T3)

Parfois, il n'y a pas de déterminant :

(121) a. L'activité de [vétérinaire rural]$_i$... (T4)
b. Les études géophysiques sur [surfaces en cours de fouille]$_i$... (T3)

Deux chaînes ne permettent pas de faire des remarques statistiques très pertinentes. On peut tout de même souligner la grande stabilité lexicale, alors que la stabilité formelle est plus faible. La « surface décapée » est parfois décrite comme « surface en cours de fouille », « sol décapé » ou bien simplement comme « sol » (du moins avons-nous considéré qu'il s'agissait du même référent).

Les catégories principales sont SN définis et SN sans déterminant, comme le montrent les exemples cités plus haut. Il n'y a aucun pronom, mais quelques déterminants possessifs (« leurs [= celles du vétérinaire] fonctions »).

Ces deux référents sont presque exclusivement présents dans l'introduction et la discussion, encore que ce soit moins vrai de la surface décapée, dont la présence est assez régulière tout au long du texte.

3.7 Les noms abstraits

Nous appelons « noms abstraits » les noms qui semblent désigner des « Idées platoniciennes », comma la « justice », « la laïcité », ou, pour ce qui nous occupe, « l'interdisciplinarité », « l'entreprenariat », « l'innovation » ou encore « le burnout ». Nous avons aussi inclus des noms prédicatifs (tels que définis dans la section « Typologie des noms abstraits et des prédicats », page 27), mais seulement ceux dont le caractère prédicatif est nié par l'absence de structure argumentale (« la fouille »), ou bien par une structure constante (« la représentation socio-professionnelle de la pratique de vétérinaire rural par les étudiants vétérinaires », en abrégé « la représentation socio-professionnelle », dans le texte T4 : la structure argumentale ne varie pas au cours du texte).

Il s'agit d'une catégorie peu homogène et que nous avons contruite sur des principes référentiels qui manque sans doute de solidité (tableau 12).

On décèle une certaine corrélation entre la longueur des chaînes et la distance intermaillonnaire. On peut former un premier groupe constitué des compétences politiques (T0), l'interdisciplinarité (T2), la susceptibilité magnétique (T3), le cursus et la représentation socio-professionnelle (T4). Il s'agit de chaînes longues à la distance intermaillonnaire brève. Elles parcourent donc tout le texte, et devraient se retrouver dans la plupart des parties.

À l'autre extrême, on peut former un groupe avec l'entreprenariat (T2) et la represéntation sociale (T3). Il s'agit de chaînes très brèves (moins de 11 maillons), mais dont la distance intermaillonnaire est au contraire très grande : ce sont des référents qui apparaissent tout au long du texte, mais très rarement.

Entre ces deux extrêmes, on trouve les autres chaînes, de longueur et de distance intermaillonnaire intermédiaires (entre 8 et 43 maillons pour la longueur, entre 90 et 543 pour la distance).

Nous allons étudier les caractéristiques linguistiques de chacun de ces groupes. Le premier groupe est surtout composé de SN définis (plus de 70 % dans chaque cas, sauf pour la susceptibilité magnétique) et de SN sans déterminant (notamment pour la susceptibilité, par exemple : « la mesure de susceptibilité »). Les maillons sont pour la plupart compléments du nom, mais on trouve aussi une part de sujets (10 % en moyenne) et, pour le cursus et la représentation

Chaîne	L CR	DI	L MA	DENS	CNSF	CNSL	PAR
(T0) Burnout	32	194.39	1.78	1.88	0.93	0.93	17
(T0) Capacité À Traiter Avec Les Autres	21	270.7	5.76	1.75	0.55	0.85	12
(T0) Compétences Politiques	88	74.24	2.85	2.75	0.92	0.99	32
(T0) Demande Psychologique	22	248.33	3.23	2	0.85	0.95	11
(T0) Épuisement	23	222.86	2.87	2.3	0.82	1	10
(T0) Influence Sociale	21	283.05	2.57	1.5	0.75	0.9	14
(T0) Intuition Sociale	16	352.13	2.69	1.45	0.71	0.93	11
(T0) Milieu Organisationnel	9	798.88	3.22	1.8	0.13	0.75	5
(T0) Sincérite Apparente	16	352.2	2.19	1.45	0.64	0.93	11
(T0) Stress	14	253.39	1.36	1.75	0.92	1	8
(T0) Travail	10	278.44	2.5	1.43	0.78	1	7
(T2) Entrepreneuriat	11	621.8	2	1.57	0.8	0.9	7
(T2) Innovation	19	351.78	1.9	1.73	0.82	1	11
(T2) Interdisciplinarité	74	98.04	2	2.47	0.93	0.99	30
(T3) Champ Magnétique	11	160	5	1.57	0.4	1	7
(T3) Couche Equivalente	8	127.71	3.88	1.33	0.57	0.86	6
(T3) Fouille	8	543.57	1.13	1.14	0.86	1	7
(T3) Susceptibilité Magnétique	43	96.14	2.42	1.54	0.76	0.98	28
(T4) Cursus Universitaire	55	152.56	2.49	2.5	0.68	0.89	22
(T4) Pratique Vétérinaire	43	206.69	3.65	1.95	0.5	0.83	22
(T4) Représentation Prof.	19	430	1.9	3.8	0.18	0.73	5
(T4) Représentation Soc.	9	965.88	2.33	2.25	0.63	0.75	4
(T4) Représentation Socio-Prof.	90	90.06	4.37	2.73	0.53	0.98	33
Moyenne	28.78	311.86	2.79	1.94	0.68	0.92	13.91

Tab. 12: Statistiques pour les noms abstraits.

socio-professionnelle (T4), une part importante (42 % et 23 % respectivement) de circonstants (par exemple « dans le cursus », « dans la représentation »). La répartion entre les parties est présentée dans le tableau 13.

parties	nb de maillons
introduction	107
methodologie	28
resultats	77
discussion	102
conclusion	23

Tab. 13: Répartitions des maillons des chaînes de noms abstraits du « premier groupe ».

On voit que ces chaînes sont surtout présentes dans l'introduction et la discussion, et dans une moindre mesure dans les résultats, sans doute parce que ce sont des notions-clés (sinon *la* notion-clé) de chacune des études.

On remarque également que la stabilité formelle est très grande, mais c'est une caractéristique de toutes les chaînes de cette catégorie, et pas seulement de celles de ce sous-groupe.

La deuxième groupe, celui formé de chaînes brèves mais très étendues, ne représente que peu de chaînes (deux). Il n'y a aucun pronom, ce qui se comprend facilement au vu de la distance intermaillonnaire. Au niveau de la distribution, l'entreprenariat se trouve dans l'introduction et la conclusion, l'autre chaîne dans l'introduction et la discussion. Les autres paramètres ne semblent pas discriminants, mais c'est problablement dû à la faible quantité de données disponibles.

Les maillons des chaînes du troisième groupe, enfin, se composent surtout de SN définis ou sans déterminants (entre 80 % et 100 % dans chaque cas). La présence ou l'absence d'expansions est très stable, ce qui est corrélé avec la stabilité formelle. Les fonctions sont très hétérogènes : cela résulte peut-être de l'hétérogénéité de cette catégorie de chaînes ; peut-être faudrait-il trouver d'autres sous-groupes. Globalement, la répartition est la suivante, donc une grande présence dans l'introduction et les résultats (tableau 14), un peu comme les chaînes du premier groupe.

partie	nb de maillons
introduction	113
méthodologie	27
résultats	82
discussion	45
conclusion	13

Tab. 14: Répartitions des maillons des chaînes de noms abstraits du « troisième groupe ».

En conclusion, il faut dire que cette catégorie de chaînes, qui ne présente pas de difficultés particulières d'annotation, peut se diviser en plusieurs sous-types, en fonction de la longueur et de la distance intermaillonnaire. Les propriétés linguistiques, lorsqu'elles sont homogènes, indiquent plutôt une forte proportion de SN définis et de SN sans déterminant.

3.8 Les noms prédicatifs

Selon la définition que nous avons donnée dans la section « Typologie des noms abstraits et des prédicats », page 27, nous avons inclus dans cette catégorie les noms qui ont une structure argumentale comme en aurait un verbe, voire un adjectif. Par exemple la phrase :

(122) La modération par les compétences politiques des facteurs de stress au travail sur la santé psychologique. (T0)

peut être mise en relation avec une autre phrase du même texte :

(123) Les compétences politiques modèrent l'impact des facteurs de stress au travail sur la santé psychologique. (T0)

Dans (122), en comparant avec (123), « modération » prend différents arguments, notamment un sujet (les compétences politiques) et un objet (les facteurs de stress). De plus, le terme régit d'autres compléments : le travail et la santé psychologique.

Nous avions proposé, à la suite Longo et Todirascu (2014), d'inclure les verbes dans les chaînes de référence. Nous ne l'avons pas fait dans cette étude préliminaire, d'abord parce que le texte qui s'y prêtait le mieux (T0) a été annoté avant que nous réfléchions à cette question, et ensuite parce qu'il conviendrait de changer quelque peu le schéma d'annotation afin d'inclure une propriété pour indiquer le type de maillon (nominal, verbal, voire adjectival), afin de pouvoir faire des calculs soit en tenant compte de ces maillons « verbaux », soit en n'en tenant pas compte.

Il y aurait par ailleurs un autre problème : comme annoterait-on ces maillons verbaux ? Si on inclut toute la structure argumentale dans l'annotation, nous inclurions toute la phrase... Ce sont des problèmes que nous n'avons pas encore résolus. Pour l'annotation des noms prédicatifs, nous avons choisi d'inclure tous les arguments, et de leur donner systématiquement pour fonction « complément du nom ». Ainsi, l'expression « par les compétences politiques » de l'exemple (122) a été considérée comme un complément du nom, et non comme un sujet ou un complément d'agent.

Comme nous l'avions évoqué au chapitre 1, les noms prédicatifs posent le problème de l'identité référentielle. Nous avons ici choisi de respecter la visée communicative des auteurs. Ainsi avons nous opposé la « consommation de protéines animales » et la « consommation de protéines végétales », parce que ces deux notions s'opposent dans l'article T1, mais nous avons inclus dans la première chaîne « la consommation de viande de jeunes animaux non encore sevrés », « la consommation de viande de cochon ou celle de cerf » ou encore « la consommation des ressources marines », parce que ces éléments ne sont pas en opposition dans l'économie générale de l'article (même s'ils le sont au niveau du paragraphe où ses expressions apparaissent). Nous avons par ailleurs choisi de créer une chaîne séparée pour « la consommation » en général, qui inclut la consommation de protéines animales et celle de protéines végétales.

Le tableau 15 montre les différentes chaînes annotées.

Nous pouvons isoler quatre chaînes : la consommation (T1), comprise comme consommation de végétaux et de viande, la consommation de végétaux (T1), l'aversion pour l'interdisciplina-

Chaîne	L CR	DI	L MA	DENS	CNSF	CNSL	PAR
(T0) Modération	35	176.97	8.69	2.06	0.12	0.82	17
(T1) Comportement Alimentaire	16	374.93	6.75	1.78	0.08	0.62	9
(T1) Consommation	5	535.25	12.4	1.25	0	1	4
(T1) Consommation Protéines Animales	18	336.65	9.39	2.25	0.06	0.94	8
(T1) Consommation Protéines Végétales	3	266.5	9	1.5	0	1	2
(T2) Aversion Pour Interdisciplinarité	3	131	5.33	1.5	0.5	1	2
(T2) Comportement Des Étudiants	8	914.14	6	1.33	0	0.83	6
(T2) Propension Pour Interdisciplinarité	23	300.14	6.78	1.77	0.36	0.82	13
(T3) Aimantation	19	182	3.58	2.11	0.17	0.94	9
Moyenne	14.44	357.51	7.55	1.73	0.14	0.89	7.78

TAB. 15: Statistiques pour les noms prédicatifs.

rité (T2) et le comportement des étudiants (T2). Ce sont des chaînes très brèves, à la distance intermaillonnaire très grande (sauf pour l'aversion, qui ne couvre qu'un tout petit passage du texte : nous l'avons annoté en contre-point de la propension à l'interdisciplinarité). Les autres chaînes sont plus significatives.

Ces chaînes sont intéressantes parce qu'elles s'opposent sur bien des aspects aux autres. D'abord, les maillons sont très longs : ils contiennent en moyenne 7,55 tokens, alors que la moyenne pour les maillons de toutes les chaînes s'établit à 3,47 tokens. Cela s'explique par la présence de la structure argumentale dans le corps du maillon.

Cette présence explique aussi la très faible stabilité formelle, alors que la stabilité lexicale est bien plus forte : ce sont surtout les arguments qui varient, et non la tête syntaxique du maillon (*i.e.* le référent). La structure argumentale peut varier de différentes manières :

- les arguments peuvent ne pas être exprimés,
- certains peuvent l'être, d'autre non, à tour de rôle,
- les arguments peuvent prendre des formes différentes (synonymes),
- les arguments peuvent changer (dans ce cas, soit on considère que le nom prédicatif renvoie à un autre référent, soit on considère que l'argument qui change n'est pas significatif dans le calcul de la référence, comme cela a été expliqué au chapitre 1).

Du point de vue de la catégorie grammaticale, ces chaînes s'opposent également aux autres. Si, comme les autres, elles ont majoritairement des maillons définis, elles ont, contrairement aux autres, également des maillons indéfinis, par exemple :

(124) Les résultats indiquent [un effet modérateur significatif des compétences politiques]$_i$ (T0).

On considère généralement que les indéfinis servent à introduire un nouveau référent (Chastain, 1975, Charolles, 2002), or le maillon de (124) n'introduit pas son référent, puisqu'il est au milieu de la chaîne. On pourrait cependant envisager qu'il introduit effectivement un nouveau référent : il initierait alors une autre chaîne. Ce nouveau référent serait l'*effet modérateur significatif*, par opposition à l'*effet modérateur* tout court ; l'ajout de l'épithète en ferait un nouveau référent. Cette approche serait plus cohérente avec ce que l'on sait de l'indéfini et de sa place dans les chaînes de référence. Cela révèle encore la difficulté de l'individuation et de l'identification des référents abstraits.

Sans surprise, l'immense majorité (entre 90 et 100 %) des maillons sont expansés (avec la structure argumentale).

Les fonctions sont majoritairement sujet et complément de verbe, ce qui est, là encore, une différence par rapport aux autres chaînes. Les deux exceptions notables sont le comportement alimentaire des sujets des Treilles (T1) et l'aimentation (T3), qui ont plus de 50 % de compléments du nom ; peut-être que ces deux chaînes seraient mieux placées dans le groupe de ce que nous avons appelé « les noms abstraits », puisque leur structure argumentale ne varie pas. Cela montrerait qu'il est possible, sur la base d'indices grammaticaux tels que la fonction, de déterminer l'appartenance de certaines chaînes aux catégories que nous sommes en train de décrire. Cela validerait, de fait, ces catégories. Cependant, il est trop tôt, et nous n'avons pas assez de données, pour conclure en ce sens.

La répartition de ces chaînes entre les différentes parties est présentée dans le tableau 16.

partie	nb de maillons
introduction	35
méthodologie	10
résultats	30
discussion	43
conclusion	13

Tab. 16: Répartitions des maillons des chaînes de noms prédicatifs.

Les chaînes se déroulent sur l'ensemble du texte, avec une présence moins marquée dans la méthodologie et la conclusion. Cela doit être confirmé par des données supplémentaires.

Cette rapide analyse a permis de montrer que les chaînes de noms prédicatifs ont un comportement manifestement différent des autres chaînes, tant dans la composition formelle de leurs maillons que dans leurs caractéristiques linguistiques.

Nous pensons, de plus, qu'il serait intéressant d'étudier si l'inclusion des verbes, voire des adjectifs, dans ces chaînes influencerait leurs descriptions statistiques et linguistiques dans une direction ou une autre.

3.9 Les ensembles flous

Nous désignons comme « flous » ces ensembles qui n'ont pas à proprement parler de définition intensionnelle, ce qui nous empêche les classer avec les ensembles dont il a été question plus haut (page 93). Soit l'auteur veut à dessein rester dans le flou (peut-être parce qu'il ne veut pas donner plus d'informations que nécessaire, respectant par là les maximes de Grice), comme dans :

(125) a. Les anomalies magnétiques de forte amplitude (y compris celles générées par [les vestiges archéologiques]$_i$) sont minimisées par un traitement adapté. (T3)
b. Ce panache pourrait être la traduction de [phénomènes anthropiques]$_i$. (T3)
c. Le risque de contamination par d'[autres occupations humaines]$_i$ se trouve minimisé. (T3)
d. ... qui constitue une information capitale pour apprécier l'homogénéité [des données biologiques et isotopiques]$_i$... (T1)
e. [Les végétaux consommés par le Groupe des Treilles]$_i$ sont des végétaux à photosynthèse en C3 provenant préférentiellement de milieux ouverts. (T1)

Soit il s'agit de valeurs qui sont pertinentes pour l'analyse, comme « les teneurs isotopiques en azote » et en carbone, qui sont la base de l'étude statistique du texte T1, mais ces valeurs ne sont pas définissables comme un ensemble fermé (le nombre des valeurs possibles est infini).

Le tableau 17 montre les référents que nous avons retenus.

Chaîne	L CR	DI	L MA	DENS	CNSF	CNSL	PAR
(T1) Données Anthropologiques	9	680	3.44	1.5	0	0.75	6
(T1) Isotopes Azote	49	85.71	3.06	4.45	0.84	0.93	11
(T1) Isotopes Carbone	27	158.39	3	2.7	0.83	0.92	10
(T1) Isotopes C et N	4	1086.33	3.25	1.33	0.33	0.33	3
(T1) Protéines Animales	29	203.25	4.48	3.22	0.21	0.83	9
(T1) Protéines Végétales	6	921.8	5.83	1.2	0.2	0.6	5
(T3) Activité Anthropique	23	188.64	3.44	1.44	0	0.52	16
(T3) Structures Archéologiques	12	241.36	3	1.33	0.5	0.8	9
(T4) Stages	11	812.8	2.46	2.2	0.44	0.89	5
Moyenne	18.89	486.48	3.55	2.15	0.37	0.73	8.22

Tab. 17: Statistiques pour les ensembles flous.

Il existe une corrélation inverse entre la longueur des chaînes et la distance intermaillonnaire : plus une chaîne est brève, plus la distance entre les maillons augmente. Cependant, nous pourrions aussi interpréter cette distribution comme nous l'avons fait pour les noms abstraits, et isoler d'une part les isotopes de carbone et azote (T1), les protéines végétales (T1) et les stages (T4), d'autre part les isotopes d'azote (T1), et former un troisième groupe avec le reste ; mais cela n'est pas compatible avec le sens de ces référents (par exemple, les isotopes de carbones, ceux d'azote, et enfin le référent qui mêle à la fois les isotopes de carbones et d'azote seraient dans trois sous-groupes différents).

On pourrait au contraire chercher à grouper, en faisant appel au sens, les isotopes d'azote et les isotopes de carbone, ce qui est compatible avec la distance intermaillonnaire et la stabilité, mais non avec le reste, notamment avec les caractéristiques linguistiques.

Celles-ci, justement, ne laissent transparaître aucune relation entre les chaînes de cette classe, ni aucune corrélation qui permettrait de former des sous-groupes.

La distribution selon les parties IMRaD montre une courbe en cloche qui culmine dans la partie « résultats » (tableau 18).

partie	nb de maillons
introduction	17
méthodologie	28
résultats	61
discussion	46
conclusion	14

TAB. 18: Répartitions des maillons des chaînes d'ensembles flous.

Cependant, les chaînes sont problablement trop hétérogènes pour que l'on puisse tirer une conclusion de ce tableau.

Nous devons conclure que, sans être fourre-tout, cette classe a été créée de façon *ad hoc*, et que c'est sans doute la raison pour laquelle il est difficile de trouver un comportement commun à l'ensemble de ses éléments. Il conviendra d'affiner la définition de cette classe avec des données supplémentaires.

3.10 Les variables liées et les chaînes locales

Pour finir cette étude préliminaire sur les chaînes de référence qui parcourt les articles scientifiques, nous voudrions nous intéresser aux variables liées, c'est-à-dire aux référents non-spécifiques, qui reviennent tout au long du texte. Par exemple, dans le texte T2, il est souvent question du « capitaine d'équipe ». Mais il s'agit d'une variable, puisqu'il peut s'agir de n'importe quel capitaine. La variable est dite liée quand sa valeur (*i.e.* son référent) est reprise par une autre expression (par exemple un pronom). Par exemple :

(126) [Un étudiant] se voi[t] proposer deux options. La première option est de créer une équipe, et en devenir capitaine... Une fois l'équipe créée, [le capitaine]$_1$ peut accueillir des coéquipiers en les invitant à rejoindre [son]$_1$ équipe : cette invitation se fait sur les pages profils des membres afin que [le capitaine]$_1$ puisse évaluer le parcours de formation et les compétences des étudiants qu'[il]$_1$ souhaite inviter. Une fois l'invitation lancée, l'étudiant ainsi invité peut accepter ou refuser l'invitation à travers l'interface, et rejoindre ou non l'équipe en question. La deuxième option est la candidature dans une équipe : un étudiant peut faire une demande pour rejoindre l'équipe de son choix en expliquant par écrit sa motivation. [Le capitaine de l'équipe concernée]$_2$ reçoit cette candidature, et [Ø]$_2$ choisit de l'accepter, de la refuser, ou de ne pas y apporter de réponse. (T2)

Il s'agit donc d'une succession de chaînes locales, qu'on ne peut que difficilement réunir au sein d'une seule chaîne, même pour les besoins techniques de l'annotation. Si nous l'avons fait dans cette étude exploratoire, c'est pour observer leur comportement afin de préparer l'étape suivante, qui est l'annotation par paragraphe.

On remarquera tout de même la plus grande proportions de pronoms. Alors que pour l'ensemble des chaînes annotées, on trouve 1,5 % de pronoms (ou même 1 % si on enlève les chaînes « auteur », qui sont très particulières), les chaînes de « variables liées » en ont 3 %. Pourtant, cela reste une faible proportion de pronoms par rapport, par exemple, aux textes lit-

Chaîne	L CR	DI	L MA	DENS	CNSF	CNSL	PAR
(T2) Candidature	13	225.67	4.46	2.6	0	0.78	5
(T2) Capitaine	15	210.79	2.6	2.5	0.6	0.9	6
(T2) Challenge	16	180.73	2.06	1.78	0.86	0.93	9
(T2) Équipe	33	94.94	2.36	2.75	0.58	0.96	12
(T2) Équipes Pour Un Challenge	14	22.08	3.5	3.5	0.36	0.91	4
(T2) Étudiant	14	180.69	3.29	2.8	0.13	0.75	5
Moyenne	17.5	152.48	3.05	2.65	0.42	0.87	6.83

Tab. 19: Statistiques pour les variables liées.

téraires[40], et cela correspond à l'intuition que l'on éprouve en consultant la liste des maillons : la plupart du temps, il y a répétition d'un terme lexical, et non pas le patron qu'on pourrait attendre :

(127) nom... pronom... pronom... [saut de paragraphe] nom... pronom... pronom...

La répartition est surprenante : toutes les occurrences se répartissent entre la méthodologie (59 occurrences) et les résultats (44). Néanmoins, il ne s'agit que d'un seul texte, ce qui est trop peu pour pouvoir tirer des conclusions.

L'annotation de ces chaînes devra tenir compte de leur statut ambigu : Elles sont tout à la fois indépendantes les unes des autres, et pourtant on a l'impression qu'elles doivent être reliées d'une manière ou d'une autre (voir le chapitre 6). Cela permettrait de voir s'il y a un élément (nouveau paragraphe, marqueur discursif, etc.) qui coupe la chaîne.

4 Conclusion

Nous avons essayé de répartir les chaînes des référents les plus saillants en fonction de leur caractéristiques référentielles, puisque c'est à ce niveau qu'apparaissent des problèmes d'annotation spécifiques. En confrontant ces classes de référents à cinq textes, nous avons, pour chacune des catégories, chercher à décrire les problèmes d'annotations et à proposer, sinon des solutions, du moins les choix que nous avons faits. Nous avons aussi décrit les caractéristiques globales de chaque ensemble (longueur, distance intermaillonnaire, coefficients de stabilité, catégorie et fonction grammaticales, expansion, distribution dans les parties IMRaD). Certaines de ces classes semblent ne pas avoir de caractère homogène, ce qui peut être le signe d'un défaut de définition. Il convient cependant de rappeler que ces classes ont été construites *a priori*, sur des caractères référentiels, et non *a posteriori*, sur des données statistiques.

Cependant, les variables liées et les chaînes locales nous ont rappelé qu'annoter les référents qui parcourent l'ensemble du texte n'est pas la seule manière d'étudier les chaînes de référence. Une approche consisterait à étudier les chaînes paragraphe par paragraphe. Il convien-

[40] Un rapide décompte des pronoms de quelques textes littéraires avec l'étiqueteur TreeTagger nous a donné une proportion d'environ 7 % de pronoms.

drait alors de chercher les chaînes de chaque paragraphe, et de les relier, par des annotations spécifiques (même référent, anaphore associative, etc.), à des chaînes d'autres paragraphes.

Cette approche donnerait des résultats probablement très différents. Un certain nombre des chaînes que nous avons étudiées dans cette partie disparaîtraient car elles n'apparaissent que sporadiquement, et jamais plusieurs fois dans le même paragraphe. En effet, la distance intermaillonnaire moyenne est proche de 340 tokens, ce qui signifie que, puisque la longueur moyenne d'un paragraphe est de 140 tokens, la plupart des chaînes n'ont un maillon que tous les deux paragraphes et demi. Mais d'autres caractères apparaîtraient ; il serait par exemple intéressant d'observer comment les chaînes évoluent de paragraphe en paragraphe, et si une même chaîne se comporte différemment dans différents paragraphes.

Nous nous proposons donc, dans la partie suivante, de présenter une deuxième étude exploratoire, dans laquelle nous avons annoter les chaînes de référence paragraphe par paragraphe.

Chapitre 6

Une deuxième étude exploratoire : annotation systématique des chaînes de paragraphe

1 Introduction

Pour cette deuxième étude, nous avons choisi d'annoter systématiquement toutes les chaînes de chacun des paragraphes des textes, sans les sélectionner avant (comme nous l'avions fait pour l'étude précédente). Le cadre dans lequel s'inscrit la chaîne n'est donc plus le texte, mais le paragraphe.

Cette approche présente l'avantage de mieux correspondre au traitement textuel et cognitif des chaînes. On peut en effet établir un lien fort entre découpage en paragraphes et chaînes de référence[41].

Plusieurs auteurs, dans des approches que Huang (2000) nomme « hiérarchiques » (voir Schnedecker et Landragin (2014) pour d'autres références), voient dans la « structure hiérarchique du discours » le principal facteur qui détermine les relations anaphoriques (p. 309). Ainsi :

> *at the beginning or peak of a new discourse structural unit tend to be done by a full NP, whereas subsequent mentions within the same discourse structural unit tend to be achieved by a reduced anaphoric expression. Structural units in discourse can be in the form of* e.g. *turns, paragraphs, episodes, events, and themes.*

[41] Pour d'autres approches, voir la synthèse que propose Schnedecker et Landragin, 2014, pp. 5–7.

Ariel (1990, pp. 18–19) choisit pour *structural unit* le paragraphe, et trouve que, par exemple, l'immense majorité (plus de 95 %) des pronoms ont leur antécédent dans le même paragraphe, et le plus souvent dans la phrase précédente[42].

Schnedecker (1997, 2005) a étudié plus spécifiquement la relation entre paragraphe et chaîne de référence :

> [Elle] a montré que le nom propre coïncidait notamment avec le découpage en paragraphes. En effet, « l'alinéa signale au lecteur qu'il vient de traiter une unité de sens et qu'il va passer à une unité ultérieure » (Bessonnat, 1988, p. 85), ce qui revient à dire que l'ouverture d'un nouveau paragraphe désactive le référent en cours. (Longo, 2013, p. 46).

Le découpage en paragraphes présente deux autres avantages. Nous avons vu dans les chapitres précédents (1 et 5) que les référents abstraits pouvait mieux être individués et identifiés sur de courts passages (par exemple des paragraphes), alors que cela se révélait difficile, voire impossible, sur des textes de plusieurs milliers de mots. Par ailleurs, les chaînes des paragraphes sont beaucoup plus homogènes, notamment du point de vue de la longueur et de la portée, et sont donc plus facilement comparables.

Pour cette deuxième étude exploratoire, nous avons annoté quatre (T1 à T4) des cinq textes que nous avions déjà annotés pour la première étude. Annoter les mêmes textes nous a en effet permis de faire des comparaisons entre les référents des deux études.

2 Spécificité des chaînes de référence de paragraphe

Nous appelons « chaînes de référence de paragraphe » des chaînes contenues dans les limites d'un paragraphe. Nous entendons par paragraphe « un espace de texte compris entre deux alinéas » (Bessonnat, 1988, p. 83), ce qui se matérialise, dans les textes que nous avons récupérés sur le portail `revues.org`, par une séparation (un petit espace blanc) entre deux paragraphes. Nous respectons donc le découpage de l'auteur (ou de l'éditeur ?), et nous ne l'avons modifié en aucune façon.

Techniquement, nous considérons que les inter-titres sont des paragraphes. Néanmoins, nous les avons exclus de cette étude (alors que nous les avions inclus dans l'étude précédente), parce qu'il n'y pas de chaîne à l'intérieur d'un seul inter-titre (trop court).

Il arrive qu'un même référent initie des chaînes dans différents paragraphes : dans ce cas, chacune des chaînes initiées compte comme une chaîne séparée, indépendante des autres. Nous introduirons plus tard ce que nous avons appelé des « chaînes partagées », terme qui permet de rendre compte des liens qui peuvent unir les chaînes initiées dans différents paragraphes par un même référent, et qui s'opposent aux « chaînes uniques », dont le référent n'initie qu'une seule chaîne de paragraphe dans tout le texte.

[42] Faute de temps, nous n'avons pu descendre au niveau de la phrase.

2.1 Dénombrements et statistiques

2.1.1 Pour l'ensemble du corpus et par textes

Nous comptons 211 chaînes, donc beaucoup plus que dans la précédente étude, qui n'en comprenait que 89. Cela est dû à la section des chaînes après chaque paragraphe : le nombre de maillons annotés est, lui, bien inférieur : 925 contre 1963.

Les chaînes de paragraphe se caractérisent surtout par leur brièveté, tant du point de vue du nombre de leurs maillons (4.4 maillons en moyenne, contre plus de 22 dans l'étude précédente) que de la distance intermaillonnaire (43 tokens, contre 338 précédemment).

Le tableau 20 montre une très grande différence des chaînes selon le texte. Le texte T3 n'a que 9 chaînes (et 28 maillons) alors que, à l'autre extrême, les textes T4 et T2 ont 81 et 86 chaînes respectivement (348 et 402 maillons). On pourrait déjà à ce niveau émettre l'hypothèse d'une différence de pratiques entre les sciences de la nature (T1 et T3, très peu de chaînes) et les sciences humaines (T2 et T4, beaucoup de chaînes), d'autant que cette différence dans le nombre des chaînes n'est pas corrélé à la longueur moyenne des paragraphes (qui s'établit en moyenne à 115 tokens par paragraphe) : le texte qui a le plus de chaînes a des paragraphes relativement courts (94 tokens, en-dessous, donc, de la moyenne du corpus).

La corrélation entre le nombre de chaînes et le nombre de tokens (*i.e.* la longueur du texte) est par contre beaucoup plus importante (le coefficient de corrélation linéaire[43] est de 0.91), mais cela n'explique pas la différence du nombre de chaînes : en effet, alors que le texte T2 est à peine 1.6 fois plus long que le texte T3 (c'est-à-dire que le nombre de tokens est 1.6 fois plus élevé dans le texte T2 que dans le texte T3), le nombre de chaînes (et de maillons), lui, est près de 10 fois plus élevé. Par contre, la corrélation entre le nombre de chaînes et le nombre de paragraphe est beaucoup plus faible (0.69) : cela s'explique par l'hétérogénéité du comportement des paragraphes vis-à-vis des chaînes, ce que nous verrons dans un instant.

Une autre corrélation est à noter : celle, très forte (0.99), entre le nombre de maillons et le nombre de chaînes, ce qui tend à confirmer l'homogénéité des chaînes du point de vue de leur longueur.

La distance intermaillonnaire est une autre différence entre les textes : elle est très basse pour les textes T2 et T3. Cela s'explique notamment par la présence, surtout dans le texte T2, de ce que nous avons appelé « variables liées » (référents non-spécifiques, comme dans *Prenez une banane, coupez-la en tranches*), dont nous expliquerons le fonctionnement plus bas.

La longueur moyenne des maillons est de 3.2 tokens : ce n'est pas différent des chaînes de l'étude précédente.

[43] Le coefficient de corrélation linéaire permet de mesurer le lien qui existe (ou non) entre deux séries de données (ici, par exemple, le nombre de chaînes et le nombre de tokens). Il peut être compris entre 0 et 1 : plus il se rapproche de 1, plus la corrélation est forte (c'est-à-dire, dans notre exemple, plus le nombre de tokens augmente, plus le nombre de chaîne augmente, à proportion égale). Au contraire, plus il est proche de 0, moins ce lien est évident (nombre de chaînes et nombre de tokens n'évoluent pas de la même façon). Il peut aussi être compris entre -1 et 0 : dans ce cas, la corrélation est inverse (plus le nombre de tokens augmente, plus le nombre de chaînes diminue).

	nombre de tokens	nombre de chaînes	longueur moyenne des chaînes (en maillons)	distance intermaillonnaire	coefficient normalisé de stabilité formelle	coefficient normalisé de stabilité lexicale	nombre de maillons	longueur moyenne des maillon (en tokens)	densité globale (en maillons par par.)	densité des chaînes (en chaînes par par.)	nombre total de par.	longueur moyenne des par. (en tokens)
T1	6208	35	4.2	75.75	0.48	0.77	147	3.9	4.2	1	35	177.37
T2	7259	86	4.67	27.66	0.48	0.8	402	2.72	4.67	1.12	77	94.27
T3	4410	9	3.11	28.7	0.22	0.83	28	3.09	3.11	0.16	55	80.18
T4	8733	81	4.3	47.46	0.31	0.69	348	3.45	4.3	1.25	65	134.35
corpus	26610	211	4.38	43.28	0.41	0.75	925	3.21	4.07	0.91	232	114.7
introduction	4853	42	4.71	37.49	0.43	0.67	198	2.76	5.35	1.14	37	131.16
méthodologie	4760	41	4.17	28.58	0.25	0.65	171	3.15	2.9	0.69	59	80.68
résultats	9846	55	3.8	43.11	0.43	0.81	209	3.46	2.3	0.6	91	108.2
discussion	5770	58	4.88	57.08	0.43	0.81	283	3.52	9.13	1.87	31	186.13
conclusion	1381	15	4.27	46.94	0.58	0.85	64	2.55	4.57	1.07	14	98.64

Tab. 20 : Statistiques globales de la deuxième étude exploratoire. Voir les sections « Méthode de calcul de quelques indicateurs », page 82, et « Autres indicateurs utilisés », page 84, pour le calcul des différents indicateurs. Nous avons ajouté ici la « densité des chaînes », qui est le nombre moyen de chaînes par paragraphe.

La densité, c'est-à-dire le nombre moyen de maillons (toutes chaînes confondues) par paragraphe est deux fois plus importante, et se monte à 4 maillons par paragraphe. La comparaison avec l'étude précédente n'est pas pertinente, puisque nous n'avions alors annoté que des chaînes que nous avions choisies, et non *toutes* les chaînes. Par contre, il faut noter que ce chiffre masque en fait de très importantes disparités entre les paragraphes.

En effet, seul 95 paragraphes (tous textes confondus) ont au moins une chaîne de référence, soit seulement 40 %. De plus, il y a de grandes différences dans le nombre de chaînes et de maillons par paragraphe. Ainsi, si la plupart des paragraphes n'ont qu'une ou deux chaînes (parmi ceux qui ont au moins une chaîne), certains peuvent avoir jusqu'à 9 chaînes et 42 maillons.

Cette distribution hétérogène correspond à la division entre parties IMRaD.

2.1.2 Par parties

Si la proportion de paragraphes contenant au moins une chaîne est relativement constante selon les parties (entre 40 et 50 %, sauf la méthodologie, à 32 %), comme le montre le tableau 21, la densité globale (telle que nous l'avons définie plus haut) est très variable : alors qu'il y a plus de 9 maillons par paragraphe dans la discussion, la méthodologie et les résultats en comptent moins de 3 en moyenne (voir le tableau 20). C'est en effet dans la discussion qu'on trouve le plus de chaînes, alors même que c'est la partie qui a le moins de paragraphes.

	T1	T2	T3	T4	TOTAL	nb de par	proportion
introduction	3	10	0	5	18	37	0.49
méthodologie	2	9	2	6	19	59	0.32
résultats	5	15	6	10	36	91	0.4
discussion	2	4	0	9	15	31	0.48
conclusion	1	4	0	2	7	14	0.5
TOTAL	13	42	8	32	95	232	0.41
nb de par	35	77	55	65	232		
proportion	0.37	0.55	0.15	0.49	0.41		

TAB. 21: Statistiques, par textes et par parties, de la deuxième étude exploratoire.

On remarquera d'ailleurs que les paragraphes de la discussion sont, par rapport à ceux des autres parties, les plus longs (186 tokens en moyenne). Or la densité est très fortement corrélée (avec un coefficient de 0.92) à la longueur moyenne des paragraphes, notée dans le tableau 20. Cela est moins intuitif qu'il n'y paraît au premier abord : nous ne parlons pas ici de la corrélation entre le nombre de maillons (ou de chaînes) et le nombre de tokens—qui est d'ailleurs beaucoup plus faible (seulement 0.66) que si l'on prend les valeurs par texte et non par partie, ce qui signifie qu'il y a un réel contraste entre les parties : les chaînes ne se comportent pas de la même manière dans toutes les parties—, mais nous parlons de la corrélation entre la densité des maillons et la longueur des paragraphes, ce qui signifie que plus le paragraphe est long, plus il sera peuplé (ou, pour prendre une image : plus il sera « noir de monde »).

Pourtant, les chaînes de la discussion ont paradoxalement la distance intermaillonnaire la plus grande (57 tokens), ce qui doit être mis en relation avec la longueur des paragraphes : plus la longueur moyenne du paragraphe augmente, plus la distance intermaillonnaire augmente. Nous avons là l'indication que les chaînes s'étendent sur l'intégralité de leur paragraphe.

La distance intermaillonnaire plus faible pour la partie « méthodologie » (29 tokens) est due à la forte présence de chaînes brèves (les « chaînes éphémères » que nous étudierons plus bas), notamment les variables liées, qui se concentrent dans cette partie (notamment dans le texte T2).

En croisant les données relatives aux textes avec celles relatives aux parties, on observe un phénomène intéressant : le texte T3, qui a un nombre très faible de maillons par rapport aux trois autres textes semble faire exception : il n'y a aucune chaîne dans la partie discussion, partie pourtant la plus densément peuplée dans les autres textes. Peut-être parce que cette partie est non seulement très courte (759 tokens contre 1670 pour les autres parties en moyenne), mais aussi et surtout très fragmentée (10 paragraphes, contre 3 et 6 pour les textes T1 et T2 ; seul T4 a plus de paragraphes, mais il est aussi deux fois plus long) : c'est donc une succession de petits paragraphes qui empêchent la formation des chaînes. Le texte T1 ne présente pas ces caractéristiques, de sorte que nous ne pouvons pas émettre d'hypothèse sur un comportement propre aux sciences « dures ».

Il est enfin intéressant de noter que la densité des chaînes (le nombre moyen de chaînes par paragraphe) est beaucoup plus importante dans la discussion (1.9 chaînes par paragraphe) que dans les autres parties. Puisque les chaînes sont également les plus longues (4.9 maillons) dans cette partie, nous comprenons bien, comme nous venons de le dire, la forte densité de cette partie.

Pour résumer cette analyse de la partie « discussion », nous dirons que c'est une partie qui a peu de paragraphes, mais des paragraphes longs, avec plus de chaînes qu'ailleurs, et des chaînes plus longues, ce qui fait que c'est une partie qui a une densité de maillons vraiment plus grande que les autres. De plus, nous avons vu que l'analyse, même à un niveau global, des chaînes de référence permettait nettement de distinguer les parties (au moins certaines d'entre elles) des articles de format IMRaD.

2.2 Comparaison des référents des deux études exploratoires

Nous voudrions maintenant comparer les référents que nous avons manuellement sélectionnés pour la première étude et ceux que nous avons trouvés après avoir annoté l'ensemble des chaînes de paragraphe.

Dans la première étude, nous avions choisi les référents les plus saillants à partir du titre, des mots-clés, du résumé et d'analyses de fréquences. Dans cette deuxième étude, la démarche est toute différente : nous avons annoté les chaînes que nous avons trouvées dans chaque paragraphe, sans rien choisir ni sélectionner à l'avance.

Il est donc intéressant d'étudier dans quelle mesure les référents trouvés lors de l'annotation systématique recoupent ceux que nous avons délibérément choisis. Cela revient à se deman-

der si les référents les plus saillants du texte initient—ou non—des chaînes de paragraphe. Nous avons donc cherché les correspondances dans les deux listes de référents.

Sur tous les référents trouvés dans cette deuxième étude, entre 25 et 45 % seulement correspondent à ceux qui nous avions sélectionnés lors de la première étude ; à l'exception du texte T3, pour lequel le pourcentage monte à 66 %, mais il faut le mettre un peu à part puisqu'il n' a que six référents, soit entre 3 et 7 fois moins que dans les autres textes. Inversement, beaucoup des référents sélectionnés précédemment ne se retrouvent pas ici (de même entre 25 et 45 %), à l'exception cette fois du texte T4, pour lequel tous les référents de la première étude, sauf l'auteur, se retrouvent dans celle-ci.

Ces chiffres révèlent que les chaînes de paragraphe ne sont pas un bon indicateur des référents les plus saillants sur l'ensemble du texte, puisque, d'une part, la majorité d'entre eux n'initient pas de chaînes de paragraphe, et que, d'autre part, beaucoup des référents de ces chaînes ne sont pas des thèmes du texte.

On peut cependant aller plus loin dans l'analyse et contraster les textes plutôt issus des sciences de la nature (T1 et T3) et ceux plutôt issus des sciences humaines (T2 et T4). Si on enlève les chaînes des variables liées que nous avions incluses dans la première étude mais dont le statut est ambigu et qui ne sont pas vraiment des référents saillants sur l'ensemble du texte, nous trouvons que plus de 62 % des référents du texte T2 choisis lors de l'étude précédente initient des chaînes de paragraphe, ce qui tend à rapprocher ce texte du texte T4. On pourrait donc faire l'hypothèse que dans les textes des sciences humaines les référents saillants au niveau du texte initient beaucoup plus fréquemment des chaînes de paragraphe que ceux des textes des sciences de la nature. Cette hypothèse devrait cependant être confirmée par plus de données.

2.3 Chaînes de paragraphe et thèmes de paragraphe

Puisque les référents que nous avons sélectionnés dans la première étude représentaient les thèmes principaux du texte (ou du moins ce que nous considérions comme tels) et que nous venons de voir que certains de ces référents, mais pas tous, se retrouvaient dans les chaînes de paragraphe, il est intéressant de voir quels sont les référents des autres chaînes de paragraphe. On pourrait penser que ce sont des référents importants, saillants, peut-être les thèmes principaux du paragraphe.

Nous avons examiné chacun des référents restants : il semble bien qu'au moins un certain nombre d'entre eux représentent les thèmes du paragraphe, c'est-à-dire qu'ils sont saillants au niveau du paragraphe. Par exemple, un paragraphe du texte T4 s'interroge sur la corrélation entre l'année d'étude, dans le cursus universitaire, des étudiants vétérinaires, le type de territoire (urbain, rural) d'installation envisagé après la fin des études et les représentations socio-professionnelles que les étudiants ont de leur futur métier. La chaîne « type de territoire envisagé » qui apparaît dans ce paragraphe, et seulement dans ce paragraphe, est donc l'un des thèmes du paragraphe.

Cependant, d'autres référents ne sont pas du tout en lien avec les thèmes principaux du paragraphe. C'est le cas, par exemple, dans le texte T1, de la chaîne « utilisation de la fumure »,

qui apparaît dans un paragraphe qui s'interroge sur une anomalie des valeurs de la teneur en azote de certains ossements. La discussion tourne autour des « légumineuses », et la fumure n'apparaît que comme une hypothèse aussitôt rejetée.

Il serait intéressant de savoir si l'ensemble des référents les plus saillants des paragraphes initient des chaînes, et donc si l'on peut déduire la présence ou l'absence d'une thématique dans le paragraphe à partir de la présence ou de l'absence d'une chaîne de référence y renvoyant. Cela n'est cependant pas possible avec les données dont nous disposons, puisque nous n'avons pas cherché systématiquement les thèmes de paragraphe.

Nous avons en tout cas remarqué que les chaînes de paragraphe qui renvoient à un thème particulièrement saillant apparaissent souvent dès la première phrase du texte, ce qui est cohérent avec, par exemple, l'analyse du paragraphe que fait Bessonnat (1988) : « l'ouverture du paragraphe est marqué très souvent... par une disjonction thématique ». Nous n'avons pas quantifié ce phénomène, mais il serait intéressant, peut-être, d'inclure lors d'une prochaine étude une annotation sur la position du premier maillon : cela permettrait sans doute de mieux discriminer entre les deux types de chaînes.

En effet, il est difficile d'évaluer quelle est la proportion de référents qui représentent les thèmes des paragraphes dans lesquels ils apparaissent, puisqu'il est difficile de décider si tel référent fait partie ou non de ces thèmes. Néanmoins, nous pouvons estimer, d'après notre observation, que dans la plupart des cas, les chaînes de paragraphe renvoient soit à l'un des thèmes saillants du texte (choisis lors de la première étude), soit à l'un des thèmes saillants du paragraphe.

Pour l'heure, nous pourrions peut-être tenter de mieux séparer les chaînes thématiques des autres en étudiant une donnée pour laquelle nous avons des valeurs : la distance intermaillonnaire.

2.4 Chaînes éphémères

On pourrait envisager un indicateur pour mesurer la dispersion de la chaîne à l'intérieur du paragraphe. Cet indicateur pourrait mêler la distance intermaillonnaire, le rapport de cette distance avec la taille du paragraphe, la position dans le paragraphe, la composition de la chaîne : si la chaîne contient des pronoms, on pourra faire l'hypothèse qu'il s'agit d'une chaîne locale. C'est d'ailleurs ce que semblent confirmer les chiffres : sur les 144 pronoms annotés (*i.e.* qui sont maillons d'une chaîne) du corpus, 77 appartiennent à des chaînes qui ont une distance intermaillonnaire de moins de 20 tokens, soit 30 % des chaînes. Autrement dit : 53 % des pronoms se concentrent dans les 30 % des chaînes les plus courtes. (En prenant pour seuil de la brièveté des chaînes une distance intermaillonnaire de moins de 10 tokens, on trouve que 21 % des pronoms se concentrent dans les 10 % des chaînes les plus courtes.)

Mais en l'absence d'un tel indicateur, nous nous appuierons sur la distance intermaillonnaire. On peut ainsi opposer des chaînes très éphémères, qui ne durent que l'espace de quelques mots, et des chaînes beaucoup plus longues.

Nous donnerons l'exemple du paragraphe 27 du texte T1, qui traite avant tout des légumineuses (qui entrent dans la composition des repas des hommes du néolithique). La chaîne « lé-

gumineuses » a une distance intermaillonnaire moyenne de 96 tokens, c'est-à-dire qu'entre chaque maillon, il y a en moyenne 96 tokens. Composée de cinq maillons, elle s'étend sur tout le paragraphe ($5 \times 96 = 480$ tokens environ[44]). La chaîne « utilisation de la fumure », quant à elle, a une distance intermaillonnaire moyenne de 16 tokens, c'est-à-dire qu'entre chaque maillon, il y a en moyenne 16 tokens ; elle est donc très réduite : composée de trois maillons, elle ne s'étend que sur $3 \times 16 = 48$ tokens, soit dix fois moins.

Le cas le plus extrême est celui d'une chaîne qui n'est pas plus courte que les autres, puisqu'elle a trois maillons comme beaucoup d'autres (rappelons que la longueur moyenne est de 4 maillons), mais elle a la plus petite distance intermaillonnaire :

(128) Il n'a pas été possible, pour des raisons de disponibilité, d'inclure des étudiants préparant [le concours C]$_i$, alors que [sa]$_i$ spécificité et [son]$_i$ mode de recrutement pourraient avoir un impact sur les représentations des étudiants le préparant. (T4)

Un autre cas extrême, où toute la chaîne est contenue dans le premier maillon :

(129) ...[tous les arguments [qui]$_i$ viennent d'être mentionnés et [qui]$_i$ auraient été susceptibles d'expliquer l'aversion des étudiants pour l'interdisciplinarité]$_i$... (T2)

Les plupart des chaînes de ce type se caractérisent par la présence de pronoms (notamment relatifs) ou de possessifs, comme ici. Si on analyse les catégories grammaticales des chaînes qui ont une distance intermaillonnaire de moins de 20 tokens (soit 62 chaînes, donc 30 % des chaînes), on constate que dans 56 % des cas, la chaîne commence par un SN défini. Cette configuration se retrouve surtout lorsque le référent est un ensemble (près des deux tiers des cas), généralement des participants (« les étudiants ») aux expérimentations des textes T2 et T4. Ce sont donc des chaînes qu'on retrouve dans les paragraphes précédents (ce que nous appellerons plus loin des chaînes « partagées »). Cela explique la forte présence de SN définis en première position. Les maillons suivants sont, à proportions à peu près égales, des SN définis, des pronoms relatifs, des déterminants possessifs et des pronoms personnels.

Ces résultats sont similaires si on prend les chaînes qui ont une distance intermaillonnaire de moins de 10 tokens, mais changent au-dessus du seuil des 30 tokens : la proportion de SN définis en premier maillon augmente pour atteindre 63 % avec l'ensemble des chaînes, et celle de SN définis en second maillon écrase les autres catégories grammaticales (plus de 56 % contre moins de 10 % pour les autres catégories). Il y a donc bien une opposition entre les patrons des chaînes très locales et ceux des chaînes plus larges, et on peut situer la limite à une distance intermaillonnaire de 20 tokens. Afin d'éviter la polysémie de « chaînes brèves » (chaînes qui ont peu de maillons ou qui ont une faible distance intermaillonnaire ?), nous nommerons « chaînes éphémères » les chaînes dont la longueur intermaillonnaire est de moins de 20 tokens (quelle qu'en soit la longueur en termes de nombre de maillons[45]), et nous réserverons

[44] Il faut rappeler que la distance intermaillonnaire est une moyenne, et ne peut donc pas donner la longueur exacte de la chaîne. De fait, le paragraphe 27 ne fait que 455 tokens ! Cela montre au moins que la chaîne couvre effectivement *tout* le paragraphe.

[45] Dans le cas de notre corpus, les chaînes éphémères sont aussi des chaînes brèves : il faudrait vérifier plus avant si un nombre de maillons plus important, à distance intermaillonnaire égale, change le comportement de la chaîne.

le terme de « chaînes brèves » pour les chaînes qui ont peu de maillons (quelle que soit la distance intermaillonnaire).

Enfin, il faut remarquer que la plupart de ces chaînes brèves appartiennent à la classe des « ensembles » (*i.e.* leur référent est un ensemble, voir page 93), dans 44 % des cas, et à celle des « variables liées » (voir page 108), dans 23 % des cas.

Cette opposition semble aussi apparaître lorsqu'on observe les patrons de chaîne. Nous les présenterons d'abord pour l'ensemble des chaînes du corpus.

2.5 Patrons de chaîne

Nous n'avons pas proposé de patrons de chaîne, c'est-à-dire de modélisation de la séquence des maillons selon leur catégorie grammaticale, dans la première étude, parce les chaînes étaient très longues (22 tokens en moyenne contre 4.4 dans cette deuxième étude). Or on peut se demander si la modélisation d'une chaîne de 22 maillons parcourant un texte de six à huit mille mots est pertinente, surtout qu'il y avait de grandes variations : certaines chaînes avaient trois maillons, d'autres en avaient plus de 80. De plus, nous n'avions pas assez de données pour modéliser des chaînes, même sur trois ou quatre maillons (nous n'avions que 89 chaînes, contre 211 ici).

2.5.1 Patrons des chaînes du corpus

Le premier maillon est dans 63 % des cas un SN défini. Il est généralement suivi d'un défini (56 %), puis d'un autre défini (65 %). Si bien qu'un quart des chaînes suivent ce schéma. Leur référent sont surtout des noms abstraits (43 %), qui initient souvent des chaînes très monotones, comme dans l'exemple (130) ou des ensembles (33 %) ; mais il peut aussi s'agir de la chaîne « article », comme dans l'exemple (131).

(130) Il concerne le niveau de la propension à [l'interdisciplinarité]$_i$ des étudiants... Leurs partis pris méthodologiques aurait pu se traduire par une aversion plutôt qu'une propension à [l'interdisciplinarité]$_i$... L'acceptation des autres disciplines nécessite alors de ne pas envisager [l'interdisciplinarité]$_i$...

(131) La première limite de [l'article]$_i$ concerne la population d'élèves considérée dans [l'article]$_i$... La seconde limite de [l'article]$_i$ concerne....

Un peu plus de 16 % des chaînes commencent par un SN sans déterminant. Dans ce cas, le deuxième maillon est également un SN sans déterminant dans 49 % des cas, et le troisième l'est aussi dans 70 % des cas. Dans l'ensemble 5.7 % des chaînes suivent ce schéma : le pourcentage peut paraître faible, mais cette combinaison est la deuxième plus fréquente après celle décrite précédemment. Cependant, elle concerne un type de référent bien spécifique : ce que nous avons appelés des « ensembles flous » (83 % des cas), comme dans :

(132) Une autre piste alimentaire peut être explorée, il s'agit de la consommation de viande de [jeunes animaux non encore sevrés]$_i$. En effet, les tissus de [jeunes animaux]$_i$ synthétisés pendant la période d'allaitement sont connus pour présenter des valeurs de δ15N plus élevées que celles des sujets adultes associées à une légère augmentation des valeurs de δ13C... La consommation de viande de [jeunes animaux]$_i$ pourrait-elle expliquer les valeurs légèrement plus hautes de δ15N et de δ13C de certains adultes ?

Il s'agit aussi de noms massifs, comme l'azote ou le carbone :

(133) Quant à l'analyse des signatures isotopiques en [carbone]$_i$ (δ13C), elle permet de cerner la source alimentaire du [carbone]$_i$, et par conséquence, le type d'environnement dans lequel l'homme s'approvisionne. Les sujets ayant vécu dans un environnement tempéré, fermé présentent des valeurs isotopiques en [carbone]$_i$ plus faibles (*ca.* -34 à -22 ‰) que les sujets issus d'un environnement chaud, ouvert (*ca.* -19 à -6 ‰).

Comme deuxième maillon après un SN sans déterminant, on trouve un SN défini dans 28 % des cas, suivi généralement d'un autre défini. Cette combinaison est celle de 1.9 % des chaînes. Il s'agit surtout d'ensembles :

(134) L'analyse menée sur les différents sous-groupes de [participants]$_i$... [les participants]$_i$... [l'échantillon]$_i$... (T4)

Les combinaisons qui commencent par un SN indéfini, qu'on imagine pourtant fréquentes, ne se recontrent que dans un peu moins de 10 % des cas. Un SN indéfini en premier maillon n'est quasiment jamais suivi d'un pronom, mais il est fréquemment suivi d'un SN défini des cas, et souvent encore d'un autre autre SN défini en troisième maillon. Cette combinaison n'est le patron que de 1.4 % des chaînes. Elle concerne notamment des variables liées :

(135) L'étape « Participation » est l'étape initiale d'[un challenge]$_i$. Elle permet d'avoir un aperçu de l'ensemble des équipes souhaitant participer [au challenge]$_i$. A cette étape, même les équipes non finalisées (effectif minimum requis pour [le challenge]$_i$ non atteint) sont présentes.

Ces informations sont résumées dans la figure 8.

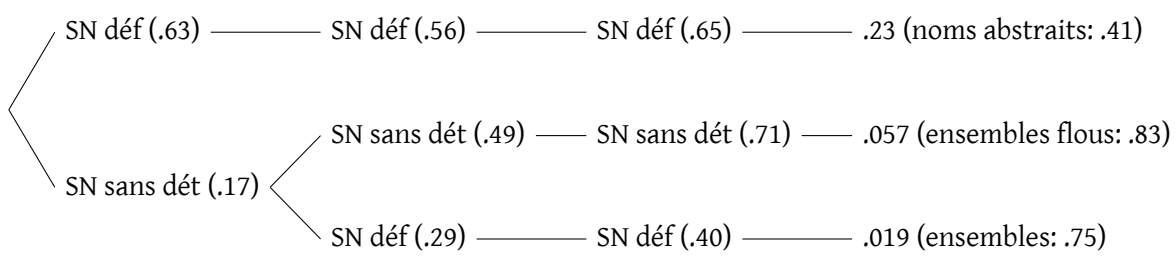

Fig. 8 : Patrons les plus courants des chaînes de paragraphe (voir page 58 pour les explications concernant une telle figure).

Nous nous sommes arrêté au troisième niveau, d'abord parce que la plupart des chaînes de paragraphe n'ont que trois ou quatre maillons, ensuite parce que nous n'avons pas assez de données pour continuer[46].

Par ailleurs, il faut noter qu'un grand nombre de chaînes ont pour patron une succession ininterrompue de SN définis. Ce sont surtout des ensembles (des participants aux études : « les étudiants »), et des noms abstraits (« l'interdisciplinarité », « le cursus vétérinaire », etc.).

D'autres patrons ne contiennent que des SN sans déterminant. Ce sont des référents caractéristiques : le nom massif « azote » ainsi que les « signatures isotopiques en carbone » et « en azote », qu'on ne trouve qu'en fonction de complément du nom.

Enfin, la chaîne « auteur » a des patrons très caractéristiques. Certains ne contiennent que des déterminants possessifs, d'autres seulement des pronoms personnels, d'autres encore un peu des deux : toutes ces combinaisons seraient impossibles pour d'autres chaînes.

2.5.2 Patrons des chaînes éphémères

Nous pouvons maintenant revenir aux chaînes éphémères, qui s'opposent aux autres chaînes non seulement par les paramètres décrits plus haut, mais aussi par leur patron.

Les patrons des chaînes dont la distance intermaillonnaire est de plus de 20 tokens sont équivalent à ceux que nous venons juste de présenter (chaînes de tout le corpus), à quelques points près. Mais les chaînes éphémères ont des patrons très différents. Le premier maillon est le plus souvent un SN défini, comme pour les chaînes plus étalées, mais aussi, dans plus de 18 % des cas, un SN indéfini (alors que pour les chaînes plus longues, c'est le SN sans déterminant qui arrive en deuxième position, avec 19 % des cas). De plus, le deuxième maillon est tout aussi souvent un SN défini qu'un pronom relatif, qu'un déterminant possessif (22 % pour chacun) ou encore qu'un pronom personnel (16 %). C'est une situation qui ne se retrouve pas avec les autres chaînes, où c'est le SN défini qui domine largement (69 % des cas), et où les pronoms relatifs, déterminants possessifs et pronoms personnels couvrent 6 % des cas ou moins. Nous ne nous aventurerons pas au-delà du deuxième maillon, par manque de données (nous n'avons que 36 chaînes éphémères).

En résumé, si l'on peut conserver, à quelques points près, l'arbre des patrons présenté plus haut pour les chaînes étalées, nous devons en composer un autre pour les chaînes éphémères (figure 9).

En étudiant les patrons, nous avons commencé, en utilisant la catégorie grammaticale, l'analyse des caractères linguistiques. Nous voudrions continuer cette analyse de façon plus globale.

[46] À ce niveau déjà les fréquences absolues sont parfois ridiculement faibles : nous proposons surtout ici une méthodologie et des hypothèses qu'il faudrait confirmer (ou infirmer) en annotant (beaucoup) plus de textes.

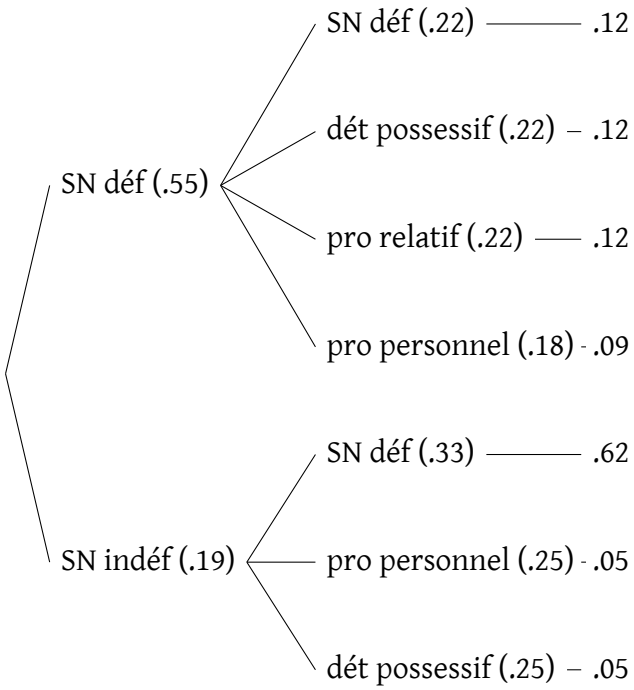

Fig. 9: Patrons les plus courants des chaînes éphémères.

2.6 Caractères linguistiques

Pour l'ensemble du corpus, l'analyse des caractères linguistiques annotés montre que près d'un maillon sur deux (49 %) est un SN défini. Ensuite vient la catégorie des SN sans déterminant (13 %, surtout les noms massifs) et des pronoms personnels (10 %, surtout l'auteur et les variables liées).

Un tiers des maillons sont expansés, essentiellement par un adjectif (45 % des expansions) ou un nom (35 %).

Près d'un maillon sur deux (45 %) est un complément du nom. L'autre moitié se partage entre sujet (23 %) et complément de verbe (16 %).

Il est difficile de voir des contrastes dans les fréquences pour chaque texte. Tout juste peut-on noter une proportion deux fois plus élevée de pronoms personnels dans le texte T2 (15 %) : cela est dû à la forte présence de la chaîne « auteur » et des variables liées ; ainsi qu'une forte proportion de SN sans déterminant dans les textes T1 (47 %) et T3 (25 %), alors qu'elle est de moins de 10 % dans les deux autres textes. Il s'agit, dans le texte T1, de deux référents : les signatures isotopiques en azote et en carbone, notamment parce qu'elles sont abrégées par un symbole : « δ15N » et « δ13C » (par exemple : « valeurs de δ15N »). Dans le texte T2, il s'agit principalement de la « susceptibilité magnétique », comme dans : « les mesures directes de susceptibilité ».

La comparaison de ces propriétés linguistiques pour les différentes parties ne permet pas de trouver de critère discriminant. Il semblerait que les chaînes de paragraphe soient homogènes dans toutes les parties.

2.7 « Chaînes uniques » et « chaînes partagées »

Nous avons défini les chaînes de paragraphe comme des chaînes qui ne dépassent pas les limites d'un paragraphe. Un même référent peut dès lors initier *plusieurs* chaînes, pourvu qu'elles soient dans des paragraphes différents. Nous aimerions cependant ici rendre compte du lien qui existe entre les différentes chaînes de paragraphe initiées par un même référent, que nous avons appelées « chaînes partagées », et les comparer aux autres, que nous désignerons comme « chaînes uniques ».

Nous ne considérons pas une chaîne partagée comme une seule et même chaîne : nous continuons à la couper à la fin de chaque paragraphe, afin de pouvoir comparer des éléments homogènes (chaînes partagées et chaînes uniques). Mais nous leur ajoutons une propriété, afin de pouvoir les répartir en deux groupes que l'on va essayer de contraster.

2.7.1 Comparaison statistique

Sur les 211 chaînes de paragraphe, 132, soit 63 %, sont « partagées ». Il faut garder à l'esprit que nous n'avons, dans notre corpus, que 105 référents. Parmi ces 105 référents, 79 n'initient qu'une seule chaîne de paragraphe. Mais les 26 autres initient 132 chaînes. Autrement dit, 25 % des référents initient plus de 60 % des chaînes de paragraphe. Par exemple, les référents des participants (les étudiants, les ossements, etc.) initient beaucoup de chaînes, sur toute la surface du texte ; ainsi, le référent « les étudiants » du texte T4 initie 12 chaînes de paragraphe. C'est également le cas de certains noms abstraits, qui représentent des thèmes du texte, comme « les représentations socio-professionnelles » du texte T4 qui traite justement des « représentations socio-professionnelles et [du] choix de la spécialisation » (c'est le titre de l'article). C'est également le cas de l'auteur.

		nombre de chaînes	nombre de maillons	longueur moyenne des chaînes (en maillons)	longueur moyenne des maillons (en tokens)	distance intermaillonnaire	densité globale	coefficient normalisé de stabilité formelle	coefficient normalisé de stabilité lexicale
chaînes uniques	T1	15	54	3.6	4.58	70.73	3.6	0.32	0.84
	T2	30	131	4.37	3.23	16.87	4.37	0.36	0.83
	T3	5	15	3	3.53	17.9	3	0.2	0.8
	T4	29	121	4.17	3.93	41.79	4.17	0.34	0.64
	Corpus	79	321	4.06	3.76	36.31	3.78	0.33	0.76
chaînes partagées	T1	20	93	4.65	3.4	79.52	4.65	0.6	0.72
	T2	56	271	4.84	2.46	33.44	4.84	0.55	0.78
	T3	4	13	3.25	2.54	42.21	3.25	0.25	0.88
	T4	52	227	4.37	3.18	50.62	4.37	0.3	0.71
	Corpus	132	604	4.58	2.89	47.45	4.28	0.45	0.75

Tab. 22: Statistiques pour les chaînes uniques et les chaînes partagées.

Les chaînes partagées sont donc plus nombreuses, et le nombre de maillons l'est aussi : près d'un maillon sur deux appartient à une chaîne partagée (tableau 22). Ces maillons sont plus courts d'environ un token (2.9 tokens contre 3.8 pour les chaînes uniques) ; cela s'explique probablement, entre autres, par une plus forte présence (d'environ 6 %) de noms sans déterminant (ce qui enlève mécaniquement un token), par exemple avec les noms massifs « azote » et « carbone », dans les chaînes partagées.

Ces chaînes ont une distance intermaillonnaire plus longue d'environ 10 tokens. Cela s'explique notamment par la présence, dans le texte T2, de nombreuses variables liées, qui sont toutes des chaînes uniques. Cela explique pourquoi il y a une telle différence entre le texte T2 et les textes T1 et T4 en ce qui concerne la distance intermaillonnaire des chaînes uniques (tableau 22), alors qu'il n'y a aucune différence significative en ce qui concerne la distance intermaillonnaire des chaînes partagées (tableau 22). Quant au texte T3, qui a une distance intermaillonnaire faible pour les chaînes uniques, il ne contient que 5 chaînes uniques, qui sont essentiellement ce que nous avons appelé des référents définis, qui semblent partager, dans ce texte (mais pas dans les autres), les caractéristiques des variables liées.

Les autres indicateurs semblent ne pas être des critères discriminants entre les deux types de chaînes.

2.7.2 Comparaison linguistique

Si l'on s'intéresse maintenant aux caractères linguistiques annotés, on remarque que les maillons des chaînes uniques sont plus souvent des SN démonstratifs (10 % contre 4 % pour les chaînes partagées). Ces SN démonstratifs se trouvent parfois en premier maillon. C'est le cas de la seule anaphore résomptive du corpus :

(136) [Cette démarche]$_i$ présenterait les inconvénients habituels des traitements statistiques réalisés sur données déclaratives. [Elle]$_i$ compléterait néanmoins le présent article et [Ø]$_i$ pourrait ouvrir des pistes nouvelle... (T2)

C'est aussi le cas où le « premier » maillon (non annoté dans cette étude) est un verbe :

(137) Le capitaine peut accueillir des coéquipiers en les *invitant* à rejoindre son équipe : [cette invitation]$_i$ (T2)

Mais ces SN démonstratifs sont, le plus souvent, de façon plus attendue, en deuxième maillon après un SN indéfini.

Les maillons des chaînes uniques sont aussi plus souvent pronoms relatifs, ce qui s'explique par la plus grande présence, comme expansion de ces maillons, de propositions relatives. En effet, la proposition relative d'un maillon est toujours introduite par un pronom qui est un maillon de la même chaîne (« [le petit chat [que]$_i$ j'ai adopté]$_i$ »), donc les deux chiffres sont liés. La plus faible présence de relatives dans les chaînes partagées peut s'expliquer, peut-être, par le fait que ces référents ont besoin de moins de précisions (en admettant que le rôle de la relative, notamment explicative, soit d'apporter des précisions), puisqu'ils sont plus saillants au niveau du texte. Pour développer cette hypothèse, il faudrait faire la distinction entre les relatives explicatives (qui sont de vraies expansions) et les relatives déterminatives (qui par-

ticipent au calcul de la référence et ne sont pas de vraies expansions), distinction que nous n'avons pas annotée. De façon générale, ce sont d'ailleurs les chaînes partagées qui ont le plus d'expansions (tous types confondus).

Pour les deux types de chaînes, la fonction la plus fréquente est celle de complément du nom. Mais pour les chaînes partagées, cette fonction se rencontre dans plus d'un maillon sur deux, contre moins d'un maillon sur trois pour les chaînes uniques. Cela s'explique par la forte présence de noms massifs ou abstraits tels que « carbone », « azote », « susceptibilité magnétique », « isotopes de carbones » et « d'azote », etc., qui sont presque exclusivement compléments du nom (par exemple « la mesure de la susceptibilité » (T3)).

Les autres fonctions fréquentes des chaînes uniques sont sujets et compléments de verbe (dans un quart des cas à chaque fois).

2.7.3 Comparaison entre les parties

L'analyse des parties révèle un contraste dans l'usage : alors que les chaînes uniques sont plus longues dans la partie « méthodologie » (4.7 tokens) que dans les parties « résultats » et « discussion » (3.7 tokens), le rapport est inversé pour les chaînes partagées, qui sont plus présentent dans la discussion (5.4 tokens) que dans les deux autres parties (3.6 tokens en moyenne). Si on considère que la longueur d'une chaîne est un marqueur de son importance, cette différence pourrait être un révélateur des différentes fonctions des parties « méthodologie » et « discussion » : alors que la première présente des référents qui n'ont pas vocation à se retrouver ailleurs (c'est le cas pour les sciences « dures », où les matériaux et outils ne sont plus repris par la suite, mais aussi pour les sciences sociales, où les outils sont des questionnaires et des tests statistiques), la deuxième reprend des référents souvent déjà mentionnés dans l'introduction, et qui le seront encore dans la conclusion.

Le nombre de chaînes dans chaque partie vient confirmer cette hypothèse : alors que les chaînes uniques sont surtout présentes dans la méthodologie et dans les résultats, les chaînes partagées peuplent surtout l'introduction et la discussion. Le nombre moyen de chaînes par paragraphe le confirme aussi, et révèle en plus une grande concentration de chaînes partagées dans la conclusion (1.9 chaînes partagées en moyenne, contre 0.2 chaîne unique).

2.7.4 Patrons

Les patrons des chaînes partagées correspondent à ceux des chaînes de l'ensemble du corpus, présentés plus haut (page 120). Ceux des chaînes uniques, s'ils sont le plus souvent :

(138) SN déf... SN déf... SN déf...

comme les autres (voir la section « Patrons des chaînes du corpus », page 120), portent la marque des patrons des variables liées, qui sont toutes des chaînes uniques, analysées ci-dessous, et l'on trouve donc souvent (un quart des cas) un SN indéfini en premier maillon, suivi dans 37 % des cas d'un SN défini (nous n'allons pas plus loin, par manque de données).

2.7.5 Îlots

Jusqu'à présent, nous avons analysé les chaînes de paragraphe, même partagées, sans nous soucier de leur environnement. Et pourtant, certaines chaînes forment ce que nous avons appelé des « îlots » : des regroupements de paragraphes *successifs* contenant les mêmes chaînes partagées. Par exemple, si un référent A initie deux chaînes dans deux paragraphes successifs, et une troisième chaîne dans un paragraphe un peu plus loin, il y a aura deux îlots.

Les chiffres-clés qui rendent compte de l'usage des îlots est la longueur de l'îlot (en paragraphes), et la différence entre le nombre d'îlots et le nombre de paragraphes dans lesquels une chaîne partagée apparaît. Ainsi, si une chaîne apparaît dans 4 paragraphes mais seulement 3 îlots, on en déduit que 3 des 4 paragraphes où la chaîne apparaît sont consécutifs. Cela peut être une indication de la saillance et de l'importance d'un référent. De plus, si une chaîne partagée est concentrée dans un seul îlot, cela signifie qu'il s'agit d'une chaîne plus locale que si elle apparaît dans trois îlots différents. On peut également calculer les distances inter-îlots, comme on calcule les distances intermaillonnaires.

Il s'agit donc d'un indicateur qui nous semble important, bien que nous ne soyons pas allé très loin dans son utilisation. Nous nous sommes contenté, pour l'instant, de l'employer pour contraster les parties sur l'ensemble du corpus.

Alors que l'introduction, la méthodologie et la conclusion ont une longueur d'îlots de moins de 1.25 paragraphes, la discussion a une longueur d'îlots de près de deux paragraphes : cela signifie que, en moyenne, un référent initiera une chaîne dans *deux* paragraphes consécutifs dans la discussion. Cela pourrait être révélateur de l'enchaînement des paragraphes dans cette partie.

Par ailleurs, nous avons remarqué que presque toutes les chaînes partagées avaient plus de deux îlots, sauf trois chaînes du texte T2 (qui sont chacune dans un îlot de deux paragraphes). Cela signifie qu'il ne s'agit pas seulement de référents qui « débordent » sur plusieurs paragraphes, mais bien de référents qui apparaissent en divers endroits, séparées par plusieurs paragraphes. La distance inter-îlot moyenne est de 11 paragraphes, mais il y a de fortes disparités entre les textes : ainsi, le texte T1 a une distance inter-îlot de 4.7 paragraphes, alors qu'elle est de 18 paragraphes pour le texte T2.

Plus révélateur est le nombre d'îlots moyen par chaîne, qui est généralement inférieur de 1 par rapport au nombre moyen de paragraphes par chaîne. Cela signifie que les chaînes partagées sont finalement peu concentrées en paragraphes consécutifs.

2.7.6 Comparaison des référents

Nous aimerions maintenant comparer les référents qui initient des chaînes partagées à ceux qui n'initient que des chaînes uniques.

Le tableau 23 révèle la présence de chaînes partagées dans les différentes classes que nous avons établies lors de la première étude.

classe	chaînes partagées	total	proportion
auteur	1	2	50 %
noms abstraits	7	13	54 %
EN et référents définis	3	11	27 %
ensembles définis	9	28	32 %
ensembles flous	2	12	17 %
référents génériques	1	5	20 %
variables liées	0	21	0 %
noms massifs	2	2	100 %
noms prédicatifs	1	9	11 %
recherche/article	2	3	67 %
TOTAL	28	106	26 %

Tab. 23: Répartition des chaînes partagées dans les classes de référents.

Nous allons reprendre ces classes une à une pour essayer de voir la différence entre les référents.

Environ la moitié des chaînes des référents abstraits s'étendent sur plusieurs paragraphes : ce sont celles qui renvoient à des thèmes de textes (par exemple la susceptibilité magnétique dans un texte (T3) sur l'emploi de la mesure magnétique comme technique de découverte archéologique, ou l'interdisciplinarité dans un texte qui veut justement tester si elle est un avantage (ou non) pour les étudiants). Les autres référents abstraits, qui ne sont pas des thématiques des textes, initient des chaînes uniques.

Ce que nous avons nommé les « ensembles flous » dans la première étude, par exemple « les jeunes animaux non encore sevrés » (T1), « les ressources alimentaires d'eau douce » (T1), « les entrepreneurs associés » (T2), etc. ne s'étalent que très rarement sur plusieurs paragraphes (2 chaînes sur 12). Ce sont en effet souvent des ensembles précis, bien que d'extension floue, dont les auteurs ont besoin très localement. Les deux exceptions sont les signatures isotopiques en azote ou en carbone (T1), qui, bien que leurs référents soient flous, apparaissent tout au long du texte, puisque c'est l'étude de ces signatures qui est la méthodologie employée dans l'article.

Les chaînes « auteur » et « article » ne s'étendent sur plusieurs paragraphes que dans le texte T2, texte dans lequel on avait vu au cours de la première étude la forte présence de ces chaînes. Dans les autres textes, l'auteur et la recherche ne parviennent pas initier des chaînes au niveau du paragraphe (un seul paragraphe du texte T1 contient une chaîne « auteur », et un seul paragraphe du texte T4 contient une chaîne « recherche »). Mais ces textes ne contiennent que peu d'expressions référant à l'auteur ou à l'article : il n'est donc pas anormal de ne pas y trouver de chaînes à ce niveau.

Les ensembles bien définis initient le plus souvent des chaînes uniques, mais il y en a tout de même une forte proportion (30 %) qui s'étendent sur plusieurs paragraphes. Ce sont d'abord les chaînes des participants aux études (les restes humains du texte T1, les étudiants du texte T2, etc.).

Les noms prédicatifs initient tous (huit cas, à l'exception de la « propension pour l'interdisciplinarité », thème du texte T2), des chaînes uniques.

Nous aurions pensé que les référents génériques initient des chaînes longues : ce n'est pas le cas. Tous (cinq sauf un : le vétérinaire rural générique du texte T4) initient une chaîne unique.

Les 21 chaînes qui renvoient à des variables liées sont, sans surprises, des chaînes uniques.

Les entités nommées et les référents définis sont un cas un peu particulier. Nous les avions rencontrés lors de la première étude : ils initiaient des chaînes avec peu de maillons, mais dont la distance intermaillonnaire était très grande. Si grande qu'ils n'apparaissaient que très rarement dans le même paragraphe. C'est pourquoi nous ne les retrouvons que très peu ici : seuls trois ont survécu, tous des entités nommées (le Groupe de Treilles et la Grotte I des Treilles du texte T1, et la plateforme Studyka du texte T2). Cependant, une autre catégorie de référents définis et d'entités nommées apparaît ici : ce sont des éléments censés être connus du lecteur, mais qui ne sont cités que pour un paragraphe, comme la salle B de la Grotte I des Treilles, le MS2D de Bartington (un appareil de mesure scientifique), ou encore le concours C pour entrer en école vétérinaire.

Enfin, les noms massifs s'étendent tous sur plusieurs paragraphes. Mais il n'y en a ici que deux : l'azote et le carbone du texte T1. Le collagène et la matière organique des textes T1 et T3 n'initient pas de chaîne de paragraphe.

Pour résumer, nous dresserons le tableau suivant :

- les noms abstraits sont partagés s'ils sont un référent saillant du texte, sinon ils ne le sont pas,
- l'auteur et l'article ne sont généralement pas partagés, tout comme les noms prédicatifs, les référents génériques, les variables liées,
- les ensembles et les noms définis sont le plus souvent uniques, mais pas toujours.

Pour analyser les référents des chaînes partagées, nous venons de reprendre les classes construites lors de la première étude. Nous voudrions maintenant étudier ces classes plus avant, et voir si les chaînes de paragraphe permettent de les opposer comme les chaînes de la première étude le permettaient.

3 Critères discriminants des classes de référents et de parties IMRaD

3.1 Problèmes d'annotation

L'annotation du corpus pour la première étude nous a posé de nombreux problèmes, que nous avons décrits dans la partie précédente. L'annotation des quatre textes pour cette deuxième étude exploratoire nous a posé bien moins de problèmes ; cela confirme la conclusion que nous avions donnée à la fin du chapitre « Le problème des référents abstraits » : l'individuation et l'identification des référents abstraits, c'est-à-dire la décision de faire renvoyer deux expressions référentielles au *même* référent, est bien moins complexe lorsque l'on est au niveau du

paragraphe. En effet, celui-ci est une unité régie par un thème et quelques référents saillants, et il est souvent facile de dire si une expression renvoie à l'un d'eux, et si c'est le cas, auquel.

La visée du locuteur est par ailleurs beaucoup plus claire dans l'espace du paragraphe, ce qui résout assez facilement les dilemmes tels que celui de savoir s'il faut inclure les légumineuses dans les végétaux (texte T1) : dans le paragraphe où il en est question, légumineuses et végétaux sont clairement en opposition (même si les légumineuses *sont* des végétaux) ; il suffit donc de créer deux chaînes. Si ailleurs les deux éléments ne sont plus en opposition, il suffit de les inclure dans la même chaîne.

Nous avons tout de même tenté de rendre compte des différentes chaînes de paragraphe lorsqu'elles renvoyaient au même référent, ce que nous avons appelé des « chaînes partagées », par opposition aux « chaînes uniques » (dont le référent n'initie une chaîne que dans un seul paragraphe). Dans ce cas, nous aurions pu rencontrer des problèmes similaires à ceux que nous avions rencontrés lors de la première étude : cela n'a pas été le cas. Le référent (surtout abstrait) d'une chaîne est une construction intellectuelle et cognitive ; et il est peut-être plus simple de relier des constructions de ce genre entre-elles plutôt que des expressions linguistiques.

Par ailleurs, l'annotation des variables liées, que nous avons annoncées à la fin de l'étude antérieure, est très facile lorsqu'on ne se souci que d'un paragraphe.

En somme, mise à part les difficultés inhérentes à la chaîne « auteur » (qui est le référent de « nous » ? quel est le statut « personnel » du pronom « on » ?), nous n'avons pas rencontré de difficultés aussi complexes que celles auxquelles nous avons été confrontées auparavant. Nous nuancerons tout de même notre propos : nous avons annoté les *mêmes* textes, et comme certains des référents sont communs aux deux études, il est probable que si certains problèmes ne se sont pas présentés ici, c'est parce que nous les avions déjà résolus lors de la première étude...

3.2 Opposition entre les classes

3.2.1 Propriétés statistiques

Nous commencerons par quelques remarques sur les différences entre les classes, sans nous soucier pour l'heure de distinction entre les textes ou les parties.

Il faut d'abord noter la très grande disparité dans le nombre de chaînes de paragraphe selon les classes (voir tableau 24). S'il y a en moyenne 20 chaînes par classe, il n'y a que deux chaînes de noms massifs (soit 1 % des 211 chaînes), mais 68 chaînes d'ensembles (soit 32 %) et 51 chaînes de noms abstraits (soit 24 %). Cela signifie que près d'une chaîne de paragraphe sur deux a pour référent un ensemble ou un nom abstrait.

Ensuite, le coefficient de stabilité lexicale est toujours supérieur à 0.5 (en moyenne : 0.67), mais s'il a une valeur de 0.5 pour les référents génériques, il est de 1 pour les noms massifs. Cela s'explique parce qu'il n'y en a que deux : l'azote et le carbone (texte T1), toujours utilisés comme complément du nom sans déterminant : « signatures isotopiques en azote », « les

	chaînes	maillons	DI	L MA	CNSF	CNSL
auteur	8	35	60.589	4.375	0.714	0.714
noms abstraits	51	226	51.688	4.431	0.426	0.785
EN et référents définis	19	69	46.665	3.632	0.328	0.57
ensembles définis	68	313	37.062	4.603	0.381	0.77
ensembles flous	20	91	44.758	4.55	0.563	0.793
référents génériques	6	28	61.079	4.667	0.195	0.5
variables liées	20	91	18.292	4.55	0.37	0.786
noms massifs	2	8	21.875	4	1	1
noms prédicatifs	11	42	56.561	3.818	0.091	0.812
recherche/article	6	22	51.833	3.667	0.542	0.708

Tab. 24 : Statistiques des chaînes de paragraphe selon la classe du référent. Voir la section « Abréviations et symboles », page v pour la signification des abréviations.

Classe de la chaîne	nb de maillons	proportion
ensembles définis	68	32 %
noms abstraits	51	24 %
ensembles flous	20	9 %
variables liées	20	9 %
EN et référents définis	19	9 %
noms prédicatifs	11	5 %
auteur	8	4 %
référents génériques	6	3 %
recherche/article	6	3 %
noms massifs	2	1 %
TOTAL	211	100 %

Tab. 25: Répartition des maillons de chaînes de paragraphe selon la classe du référent.

plantes fixatrices d'azote », etc. Ces deux référents ont d'ailleurs une longueur de maillon d'un token (« azote », « carbone »).

Concernant les noms prédicatifs, on note le même phénomène que nous avions constaté lors de la première étude : le coefficient de stabilité formelle est très bas, alors que le coefficient de stabilité lexicale est élevé : la tête du syntagme reste la même, mais les arguments varient, ou du moins n'apparaissent pas toujours, ou pas toujours dans le même ordre. Les noms prédicatifs ont, par ailleurs, les chaînes les plus longues.

3.2.2 Caractères linguistiques

L'étude des propriétés linguistiques, c'est-à-dire des caractères que nous avons annotés et qui sont décrits dans le schéma d'annotation, ne permettent pas d'opposer aussi clairement les classes que dans la première étude.

Pour l'ensemble des classes, la catégorie qui prédomine largement (souvent aux alentours de 60 %) est le SN défini, à l'exception de la chaîne « auteur », qui privilégie le pronom personnel et le déterminant possessif ; les ensembles flous qui sont majoritairement sans déterminant, les noms massifs (« azote » et « carbone »), toujours sans déterminant, et les chaînes « recherche » et « article », qui sont souvent des SN possessifs, en lien avec la chaîne « auteur » (« notre étude », « notre article »).

Les catégories secondaires ne dépassent généralement pas les 15 %, ce qui signifie qu'il y a peu de variation : les référents d'une classe donnée gardent souvent la catégorie grammaticale sur l'ensemble de la chaîne. Les exceptions sont la chaîne « auteur » (entre 40 et 50 % de pronoms personnels et de déterminants possessifs), les référents génériques (25 % des SN n'ont pas de déterminants) et les chaînes « recherche » et « article » qui alternent leurs possessifs avec un article défini dans 30 % des cas.

Le taux d'expansion tourne autour de 30 %, à l'exception, encore une fois, de la chaîne « auteur » et des noms prédicatifs, expansés dans 57 % des cas. Mais ce ne sont pas là des surprises, ni des éléments discriminants, puisque la chaîne « auteur », à cause de la nature grammaticale de ses maillons, ne peut pas avoir d'expansion, et les noms prédicatifs ont pour expansion leur structure argumentale sous-jacente. Les expansions sont dans l'extrême majorité des cas des adjectifs, sauf, encore une fois, pour les noms prédicatifs, où il s'agit de noms. Les relatives sont rares (de l'ordre des 2 % à chaque fois).

La fonction grammaticale n'offre pas plus de critère discriminant : la fonction de complément du nom est toujours majoritaire, souvent dans des proportions comprises entre 40 % et 50 %. Cela concerne également la chaîne auteur, puisque nous avons considéré que le déterminant possessif était un complément du nom, pour les raisons expliquées dans le chapitre 4. Les deux seules exceptions sont les variables liées, qui sont majoritairement (34 %) sujets (et complément d'un verbe dans 31 % des cas), et les noms prédicatifs, sujets dans 43 % des cas (et complément du nom dans 24 % des cas). Cela se comprend aisément pour les variables liées, qui sont généralement des humains de la partie « méthodologie » du texte T2, qui décrit le fonctionnement d'une plateforme Internet où des étudiants peuvent s'inscrire pour participer à un challenge. Les étudiants sont alors décrits comme des acteurs (et sont donc souvent sujets) des différentes opérations qu'ils doivent effectuer pour participer. Nous n'avons pas d'expli-

cation en ce qui concerne les noms prédicatifs, mais le même phénomène avait été noté pour ceux de la première étude (majoritairement sujets et compléments de verbe) alors même que ce ne sont pas les mêmes référents qui ont été annotés. Nous pouvons donc émettre l'hypothèse d'un comportement particulier des noms prédicatifs : outre la longueur importante de leur maillon et le nombre de leurs expansions, ils seraient plutôt sujets et objets, à la différence des autres noms. Cependant, cette hypothèse demande à être confirmée en prenant en compte l'ensemble des noms, et pas seulement ceux qui entrent dans une chaîne de référence, comme ici.

3.2.3 Patrons remarquables

Le premier maillon des variables liées est le plus souvent un SN indéfini, suivi d'un SN défini (et encore d'un SN défini ensuite, mais nous avons trop peu d'occurrences pour qu'on puisse tirer des conclusions au-delà du deuxième maillon). C'est le schéma usuel attendu :

(139) SN indéfini... SN défini... (SN défini...)

Cependant, ce schéma ne se trouve que dans 10 % des chaînes de variables liées. Le schéma le plus fréquent (25 %) est une suite de SN définis. Il s'agit soit d'anaphores associatives, soit de la reprise d'un élément de groupe :

(140) Si nous ne prenons pas en considération la formation des étudiants à l'origine de la demande d'intégration à une équipe, <u>les capitaines hommes et femmes</u> présentent... (T2)

(141) Un étudiant intéressé par un challenge peut s'y inscrire puis se voir proposer deux options... <u>La deuxième option</u> est la candidature dans une équipe... (T2)

Les deux autres classes où nous avons suffisamment d'occurrences pour tenter une analyse de ce genre, les noms abstraits et les ensembles, ne diffèrent pas beaucoup du schéma général (suite de SN définis).

3.3 Opposition entre les textes

Si l'on compare ensuite les différentes classes dans les différents textes, il faut d'abord noter que toutes les classes n'apparaissent pas dans tous les textes. Par exemple, il n'y a pas de noms abstraits dans les textes T3 et T4, ni d'ensembles flous dans le texte T3. Le cas extrême est celui des noms massifs, qu'on ne trouve que dans le texte T1.

On peut ensuite remarquer que si le texte T1 comporte essentiellement des ensembles flous, il s'agit en fait surtout de deux référents qui reviennent à plusieurs reprises (les signatures isotopiques en azote et en carbone). Cela montre l'une des limites de cette étude des chaînes de paragraphe : beaucoup d'entre elles sont des chaînes partagées. Peut-être aurait-il fallu étudier les parties en fonction des chaînes uniques seulement.

Le texte T2 comporte surtout des noms abstraits (notamment l'« interdisciplinarité »), des ensembles (notamment des participants), et des variables liées, véritables spécificités de ce texte.

Ces chaînes de variables liées sont assez nombreuses (20 chaînes) et éphémères (distance intermaillonnaire de 17 tokens).

Le texte T4 comporte surtout des noms abstraits (notamment « représentation professionnelle, socio-professionnelle, sociale ») et des ensembles (encore une fois, notamment des participants)

Les noms abstraits dans les textes T2 et T4 (12 et 34 chaînes contre 1 et 4 dans les textes T1 et T3) et les ensembles (30 et 34 contre 4 et 1) pourraient être des spécificités des textes des sciences humaines. En tout cas, leur présence permet de distinguer les deux groupes de textes.

Il ne semble pas y avoir d'autres différences significatives qui permettraient d'opposer les textes entre eux à l'aide des classes de référents.

3.4 Opposition entre les parties

3.4.1 Distribution

Le tableau 26 présente le nombre de chaînes par partie et par classe. Nous noterons que les noms abstraits, qui sont souvent les thèmes du texte (interdisciplinarité dans le texte T2, représentation socio-professionnelle dans le texte T4, etc.) se trouvent principalement dans l'introduction et la discussion, ce qui confirme la discussion ci-dessus sur les chaînes partagées.

classe	introduction	méthodologie	résultats	discussion	conclusion	TOTAL
auteur	3	0	2	2	1	8
noms abstraits	14	4	9	22	2	51
EN et réf. définis	7	4	3	2	3	19
ensembles définis	10	15	23	15	5	68
ensembles flous	1	3	6	9	1	20
réf. génériques	1	0	2	3	0	6
variables liées	0	13	7	0	0	20
noms massifs	0	2	0	0	0	2
noms prédicatifs	4	0	3	4	0	11
recherche/article	2	0	0	1	3	6
TOTAL	42	41	55	58	15	211

Tab. 26: Répartition des chaînes de paragraphe selon la partie IMRaD et la classe du référent.

On remarquera aussi la très forte présence des variables liées dans la partie « méthodologie », mais comme cela concerne presque exclusivement le texte T2, nous ne pouvons pas vraiment en tirer de conclusion.

3.4.2 Propriétés statistiques et caractères linguistiques

Nous avons trop de classes et trop peu de données, si bien que les fréquences sont bien faibles et peu exploitables : bien souvent, il n'y a qu'une ou deux chaînes dans telle partie de telle

classe. Même quand il y a plus de données, il ne semble pas y avoir de critères discriminants. Nous ne pouvons donc pas opposer les parties en fonction des différentes classes.

4 Conclusion

Nous voudrions d'abord rappeler les nouveaux termes que nous avons introduits :

- Les *chaînes de paragraphe* sont des chaînes contenues dans les limites d'un paragraphe.
- Les *chaînes uniques* sont des chaînes dont le référent n'initie une chaîne que dans un seul paragraphe du texte, alors que les *chaînes partagées* sont des chaînes dont le référent initie une chaîne dans plusieurs paragraphes.
- Les *chaînes éphémères* ont peu de maillons, et leur distance intermaillonnaire est très brève (moins de 20 tokens). Elles s'opposent aux *chaînes brèves*, qui ont peu de maillons, mais dont la distance intermaillonnaire est plus grande.
- Les *îlots* sont des ensembles de plusieurs paragraphes consécutifs dans lesquels un même référent initie des chaînes de paragraphe.

Cette étude a surtout permis de caractériser et d'opposer les chaînes entre elles, sur des critères statistiques plutôt que linguistiques. En effet, nous avons vu que les chaînes de paragraphe sont relativement homogènes, mais qu'une étude de détail permet d'opposer les chaînes éphémères aux autres. Cette notion est importante lorsqu'on s'interroge sur la relation entre une chaîne et le thème du paragraphe dans lequel elle apparaît, c'est-à-dire sa saillance : une chaîne éphémère, de portée très locale, est moins susceptible d'être initiée par un référent saillant.

L'opposition entre chaînes uniques et chaînes partagées pourrait servir à distinguer les référents saillants sur l'ensemble du texte des référents saillants sur un paragraphe.

Par ailleurs, il existe une véritable opposition entre les parties IMRaD : la discussion se distingue bien des autres en ce qu'elle a une plus forte densité de chaînes et de maillons par paragraphes.

Nous avons aussi observé que la relative homogénéité des chaînes, au moins en ce qui concerne leur nombre de maillons, permettait de calculer aisément des patrons, et nous avons découvert que dans près d'un cas sur quatre, la chaîne est une succession de SN définis.

La notion d'îlot que nous avons introduite permet de séparer les chaînes partagées qui se trouvent sur tout le texte, et celles qui, simplement, « débordent » sur un ou deux paragraphes ; ce qui est un bon indicateur de la saillance de leur référent, soit sur l'ensemble du texte, soit sur un passage.

La recherche de critères discriminants des classes de référents (établies lors de la première étude) s'est révélée par contre plus limitée, peut-être parce que les classes définies pour les chaînes initiées par des référents saillants sur tout le texte ne sont pas adaptées pour l'étude des chaînes au niveau du paragraphe. Cela montre une vraie différence entre les deux études exploratoires que nous avons proprosées ; elles sont complémentaires, et permettent chacune d'étudier des phénomènes différents. Alors que la première étude (chapitre 5) a permis de créer des classes de référents et d'opposer les parties IMRaD, cette deuxième étude (ce cha-

pitre) a été plus focalisée sur l'étude des chaînes elles-mêmes et de leurs différences intrinsèques.

Conclusion

1 Bilan et apport

Notre but était de caractériser les chaînes de référence dans les articles de recherche de format IMRaD, en essayant de trouver des critères qui permettent d'opposer les différentes parties de ce format entre elles.

Ce faisant, nous nous sommes d'abord interrogé sur la notion de *référent abstrait*, puisque nous ne pouvions pas ne pas les annoter (il y a très peu de référents humains ou concrets dans les articles de recherche). Nous avons vu que cette notion est problématique, du moins à l'échelle d'un texte, car il est difficile d'individuer les entités abstraites (le taux de fécondité allemand est-il une autre entité que le taux de fécondité français ?). Quant aux noms prédicatifs, nous ne sommes pas sûr qu'ils aient une référence, et donc initient une chaîne de référence (les philosophes, et même des linguistiques tels que Kleiber, répondraient de façon négative). Nous avons tout de même pris le parti de les annoter, ce qui nous a permis de répartir les référents de nos textes en dix classes. Nous avons trouvé que certaines de ces classes s'opposaient par le comportement particulier de leurs chaînes.

Nous avons également pu établir que les chaînes de référence étaient l'un des traits linguistiques qui pouvaient servir à opposer les parties IMRaD entre elles, même si, faute de données en quantité suffisante, nous n'avons pas pu pousser l'analyse jusqu'à étudier le comportement des référents de chacune des classes dans chacune des parties IMRaD.

Notre deuxième étude exploratoire, centrée sur les chaînes limitées à un paragraphe, a permis de caractériser le comportement de certaines chaînes particulières, et de distinguer des chaînes de paragraphe, des chaînes partagées, des chaînes éphémères. Ces notions dépassent le seul cadre des articles de recherche, et pourront être réutilisées pour d'autres corpus, et d'autres types de textes.

Les deux études exploratoires que nous avons menées, l'une consistant à annoter un petit nombre de référents saillants, l'autre consistant à annoter toutes les chaînes de paragraphe, sont donc complémentaires. Elles n'ont pas la même portée, ni la même visée : la première a permis une meilleure caratérisation des référents abstraits et du format IMRaD, la seconde

une meilleure appréhension des chaînes de référence en général, et pourrait permettre une meilleure compréhension de leur contribution à la structuration textuelle.

Sur un plan plus technique, nous avons d'abord caractérisé la distribution du format IMRaD dans le paysage scientifique français en prenant pour corpus l'ensemble des revues du portail revues.org.

Ensuite, nous avons élaboré une interface d'annotation des chaînes de référence, interface qui nous a permis d'annoter des textes plus rapidement que les autres logiciels actuellement disponibles, tout en assurant la compatibilité avec deux d'entre eux (Glozz et Analec). Nous avons aussi élaboré une série de scripts d'analyses statistiques, qui, bien que basiques, nous ont permis de traiter rapidement les quelques trois mille maillons que nous avons annotés.

Enfin, nous avons établi un schéma d'annotation, fruit d'un compromis entre l'intérêt linguistique et des impératifs plus techniques (pour l'annotation ou pour les analyses statistiques).

Malgré ces apports, notre étude souffre de certaines limites.

2 Limites

2.1 La particularité des référents abstraits

La première limite concerne les référents abstraits et les prédicats. La philosophie du langage ne nous a pas donné de réponse claire sur ce que nous devions en faire. Nous les avons annotés, mais nous aurions peut-être dû nous interroger plus longuement sur la pertinence de telles annotations, d'autant que celles-ci se sont révélées extrêmement délicates. Il y a une grande part de subjectivité dans l'individuation et l'identification de ces référents (l'assassinat de César et la mort de César sont-ils un seul événement ?), et les critères linguistiques ne sont pas suffisants pour pouvoir décider avec certitude. Un autre annotateur aurait donc problablement annoté les mêmes textes de façon quelque peu différente.

Aussi, nous ne sommes pas sûr qu'annoter des référents abstraits soit une bonne idée, et nous y aurions renoncé s'il y avait eu d'autres référents (humains ou du moins bien définis) en quantité suffisante dans les articles de recherche.

2.2 Des classes construites selon des critères non homogènes

Nous avons, dans le chapitre 5, essayé de construire des classes à partir de critères relevant du sens, des problèmes d'annotation que nous avons rencontrés, des contrastes dans les indices linguistiques et des différences dans les statistiques des chaînes (par exemple leur longueur ou leur distance intermaillonnaire). Mais, ce faisant, nous avons créé des classes en fonction de critères hétérogènes. Ainsi, certains noms sont rangés dans la classe des noms prédicatifs sur des critères syntaxiques et sémantiques, alors que d'autres sont rangés dans la classe des référents génériques sur des critères référentiels. Un même nom pourrait ainsi appartenir à plusieurs classes. C'est le cas, par exemple, de la « modération des effets psychologiques du burnout » dans le texte T0 : il pourrait être considéré à la fois comme un nom prédicatif, un

nom abstrait ou encore un référent générique. Par ailleurs, le peu de variables liées que l'on trouve dans la deuxième étude ne signifie peut-être pas qu'il y en a peu, mais simplement qu'elles sont classées dans une autre catégorie (les ensembles, les référents abstraits, etc.).

Cela révèle que nos classes de référents devraient être affinées, ou même qu'il faudrait faire un classement différent pour chacun des critères considérés. Nous ne l'avons pas fait parce que cela aurait demandé de conduire chacune des deux études exploratoires pour les niveaux syntaxique, sémantique et référentiel, soit six études au total. De plus, les résultats ne seraient peut-être pas forcément plus pertinents, car chaque niveau d'analyse serait parasité par les deux autres : le nombre significativement plus élevé d'expansions pour les noms prédicatifs est-il dû à de la syntaxe (ils ont besoin d'avoir des compléments exprimés), à la sémantique (ils ont une large structure argumentale), ou à la référence (spécifique ou générique, selon le cas) de ces noms ?

Les classes de référents que nous avons définies dans la première étude exploratoire soulèvent un autre problème : nous les avons réutilisées telles quelles dans la seconde étude exploratoire, alors même que les chaînes sont très différentes. Or ces catégories ne sont peut-être pas universelles, et ne sont peut-être pas adaptées à l'étude des chaînes au niveau de chaque paragraphe.

2.3 Limites de l'annotation

Par ailleurs, notre annotation est loin d'être parfaite ; le schéma d'annotation a en effet été élaboré *en cours* d'annotation : c'est pendant l'annotation que nous avons adapté et optimisé le schéma, en fonction des problèmes rencontrés et des premiers résultats. Même si nous nous sommes efforcé de corriger les annotations déjà effectuées, il reste sans nul doute un certain nombre d'incohérences.

3 Perspectives

3.1 Compléter nos annotations

3.1.1 Marquer toutes les expressions référentielles

Dans ces études exploratoires, nous n'avons pas annoté toutes les expressions référentielles, mais seulement celles qui étaient des maillons de chaînes. Il serait cependant intéressant d'annoter toutes les expressions référentielles, même celles qui n'entrent pas dans des chaînes. Cela permettrait de faire une étude contrastive entre les référents qui initient des chaînes et ceux qui ne le font pas.

Mais cela n'irait pas sans difficultés : en plus de toutes celles que nous avons évoquées au cours de ce travail, il faudrait choisir de prendre en compte, ou non, les adverbes de lieu ou de temps (*jadis*), les déictiques (*ici*, *aujourd'hui*), etc. qui initient rarement des chaînes mais peuvent être considérées comme des expressions référentielles.

3.1.2 Marquer *toutes* les expressions référentielles

Contrairement à ce que nous avions suggéré dans le chapitre 1, nous n'avons pas annoté les prédicats verbaux ou adjectivaux, mais seulement les prédicats nominaux, ce qui nous paraît manquer de cohérence. Cependant, si on choisit de n'annoter aucun prédicat, ni nominal, ni verbal, ni adjectival, on risque d'éliminer indûment beaucoup de référents abstraits sur des critères syntaxiques ou sémantiques, et non référentiels. Si au contraire on choisit d'annoter tous les prédicats, notamment les verbes, on est confronté à des difficultés jusque-là inédites : doit-on annoter toute la phrase ? doit-on inclure les circonstants ? les marqueurs énonciatifs ?

3.2 Étudier ce que nous avons laissé de côté

3.2.1 Étudier la titraille

Nous avions évoqué dans l'introduction et au chapitre 2 les travaux de Jacques sur les intertitres, et leur possible implication dans la référence et leur influence sur les chaînes de référence. Nous avions prévu des métadonnées pour « baliser » les titres, et les traiter à part. Nous n'avons cependant pas eu le temps d'étudier toutes ces questions.

3.2.2 Approfondir l'étude des chaînes de paragraphe

Dans le chapitre 6, nous avons commencé l'étude de ce que nous avons appelé les « chaînes de paragraphe ». Il nous semble intéressant de développer cette approche, notamment en incluant une analyse au niveau de la phrase (nous avons déjà évoqué l'étude d'Ariel (1990) qui établit une corrélation entre lien anaphorique et position dans la phrase : cela pourrait servir à mieux caractériser les patrons de chaînes), mais aussi en étudiant ce qu'on peut qualifier d'un « débordement » d'une chaîne de paragraphe sur les paragraphes précédent ou suivant. En effet, comme l'indique Bessonnat (1988, pp. 89–90), le thème du paragraphe suivant est annoncé dès la fin du paragraphe courant, alors que celui du paragraphe courant est rappelé au début du paragraphe suivant ; il y a donc un chevauchement qui peut avoir une incidence sur les chaînes de référence.

3.2.3 Étudier les anaphores résomptives

Les anaphores résomptives sont fréquentes dans les textes académiques, par exemple :

(142) Le rapport Carbone/Azote montre quant à lui un coefficient de corrélation de ,71 avec la susceptibilité magnétique massique. [Cela]$_i$ suggère... (T3)

Cependant, elles ne semblent pas faire chaîne (c'est-à-dire qu'il n'y a qu'une seule reprise, et non pas trois). Mais nous n'avons pas étudié le phénomène systématiquement, ce qu'il faudrait peut-être faire, parce que si elles, ou au moins certaines, initient des chaînes, elles permettraient de caractériser la progression textuelle, comme le rappellent Lundquist, Minel et Couto (2012) : « ce type d'expression linguistique... joue un rôle central au niveau discursif

en tant que “mouvement discursif” » (terme repris de Swales (1990, 2004)). Cette dernière remarque suggère à tout le moins que les anaphores résomptives pourraient être des bornes pour les chaînes.

3.2.4 Étudier le mode de cohabitation des chaînes

Nous n’avons pas non plus étudié le mode de cohabitation de chaînes, que nous avions évoqué en introduction (voir Schnedecker et Landragin, 2014). Là aussi, c’est une étude qui reste à faire, et qui semble très prometteuse tant les chaînes des articles de recherche semblent, pour certaines, s’entre-mêler, pour d’autres, se succéder.

3.2.5 S’interroger sur la relation entre « variable liée », « anaphore liée » et « chaîne liée »

Nous avons créé une classe de référents nommée « variables liées ». Nous comprenons ce terme au sens logique et référentiel : le référent peut être n’importe quel élément d’un ensemble, comme dans :

(143) Prenez *une banane* et coupez-*la* en morceaux.

Il ne faut donc pas confondre cette notion avec celle d’« anaphore liée » ou de « chaîne liée ». Ces deux dernières notions, pourtant, nous semblent importantes pour l’étude de ce que nous avons appelé les « chaînes éphémères », c’est-à-dire des chaînes très courtes.

Le concept d’« anaphore liée » doit ici être compris dans le sens qu’il a en syntaxe générative. On dit qu’il y a anaphore liée lorsque « l’antécédent… est choisi en vertu d’un calcul purement syntaxique » (Corblin, 1995, p. 157, voir aussi les développements plus théoriques de Corblin (1985b) et Milner (1982)). Corblin donne « un exemple typique » :

(144) $[\text{Pierre}]_i$ $[\text{se}]_i$ regarde dans la glace.

Pour mieux comprendre cette notion, il convient de noter qu’il y a anaphore liée lorsque l’anaphorique ne peut être que dans *la même phrase* que son antécédent. Au contraire, l’anaphore libre, qui s’oppose à l’anaphore liée, « peut toujours renvoyer à un terme extérieur à la phrase contenant l’anaphorique » (Milner, 1978, qui nous semble donner, dans cet article, la meilleure définition de l’anaphore liée).

L’anaphore liée concerne donc, par exemple, les pronoms réfléchis et les pronoms relatifs.

Si nous avons évoqué un peu longuement ces anaphores, c’est parce qu’elles donnent lieu à ce que Corblin (1995) appelle des « chaînes liées », dont la caractéristique la plus intéressante serait que « la construction de la chaîne est exclusivement régie par la syntaxe » (p. 157). Il serait donc intéressant de pouvoir isoler ces chaînes d’un type particulier (notamment en relation avec nos « chaînes éphémères »).

Ces anaphores liées pourraient aussi poser un problème référentiel. En effet, d’après ce que nous comprenons, Charolles (2007) refuse la référence aux anaphoriques liés :

> ce n'est évidemment pas un hasard, si les anaphoriques liés, qui n'ont pas de pouvoir référentiel, sont intégrés syntaxiquement, voire assument des fonctions de connecteur syntaxique, comme c'est le cas du relatif.

Il s'agit en tout cas d'une notion qu'il faudrait étudier plus avant.

3.3 Développer la notion de *chaîne*

Nous avons insisté, dans cette recherche, sur la notion de *référence* (voir chapitre 1), mais nous n'avons pas développé la notion de *chaîne* (relation anaphorique ; relation coréférentielle ; relation entre les chaînes, par exemple, relation d'anaphore associative ; découpage en sous-chaîne ; etc.). C'est ce qu'il conviendrait de faire la suite de ce travail, notamment en s'appuyant sur Schnedecker (1997) et Corblin (1995).

3.4 Étudier l'écriture à plusieurs mains

La plupart des textes que nous avons annoté ont été écrits par plusieurs auteurs. Même si le format IMRaD était censé nous prévenir contre une trop grande variation entre les auteurs (du fait de sa grande standardisation), nous ne pouvons exclure que certaines des différences entre les parties puissent être le fait d'un changement d'auteur. C'est ce que nous apprend Milard (2007), qui évoque ici le cas des chimistes :

> Le cas peut-être le plus éloquent concerne l'organisation du texte en parties spécifiques (le plan IMRED) correspondant à un mode d'organisation du travail collectif des chercheurs. Ce plan permet en effet de hiérarchiser les opérations de rédaction qui sont pratiquées dans l'équipe. Typiquement, alors que l'étudiant en doctorat ou en DEA rédige la partie expérimentale, le post-doctorant ou le jeune chercheur rédige les parties résultats et discussion et c'est au directeur de l'équipe ou au chercheur confirmé que revient la rédaction de l'introduction et l'élaboration de la bibliographie.

Dans le même ordre d'idées, peut-être faudrait-il s'interroger sur la polyphonie du texte scientifique (par exemple Grossmann, 2010), notamment dans les passages de paraphrases ou de citations.

3.5 Comparer avec les autres genres

Nous avons évoqué en introduction l'importance de la comparaison entre les différents genres pour l'étude des chaînes de référence. Nous en sommes pour l'heure resté à comparer des parties IMRaD, éventuellement des disciplines (sciences naturelles et sciences humaines), mais nous n'avons en aucun cas comparé avec d'autres genres. La faible quantité de données ne nous a pas, en effet, permis de le faire.

3.6 Interroger les spécificités du format IMRaD

Pour l'heure, nous n'avons annoté que peu de propriétés linguistiques : seulement la catégorie, la fonction et l'expansion. C'est qu'il s'agissait surtout de voir le problème des référents abstraits, d'élaborer des outils et des techniques d'analyses, et de construire un schéma d'annotation.

Nous pourrions cependant chercher d'autres paramètres linguistiques à annoter, et nous pourrions pour cela interroger les études, très nombreuses, qui ont déjà été faites sur le format IMRaD. Par exemple, nous pourrions nous inspirer de Müller-Gjesdal (2013), qui étudie un corpus d'articles de médecine au format IMRaD et en français ; Brett (1994), qui travaille sur la partie « résultats » des articles de sociologie (en anglais) ; Williams (1999), qui travaille sur la même partie, mais dans des articles de médecine en anglais, s'interrogent par exemple sur le type (c'est-à-dire le domaine sémantique) des verbes. Il serait sans doute intéressant de s'inspirer de ces recherches et d'étudier la relation entre les chaînes, ou certains types de chaînes, et le type de verbe.

Une analyse plus appronfondie de la littérature sur le format IMRaD permettrait aussi de fournir des explications aux phénomènes que nous avons constatés ; en effet, nos observations ont pour l'heure été de l'ordre de la description, et non de l'explication. Nous pourrions pour cela utiliser la notion de « fonction rhétorique » que nous avions évoqué en introduction : chaque partie IMRaD a une fonction bien définie, Par exemple, Rinck (2010) remarque que

> les sections de résultats, de discussion, et de conclusions... doivent permettre d'interroger les épistémologies en jeu, comme l'ont montré par exemple les travaux sur la place croissante de la partie méthodologique et sa fonction dans les régimes de la preuve.

Une même partie peut se diviser en « mouvements » (Swales, 2004, p. 228), à l'image de l'introduction qui en a trois : l'un pour établir un sujet, un autre pour établir une « niche », et le dernier pour occuper cette niche (Swales, 1990). Le comportement particulier des chaînes de référence pourrait peut-être s'expliquer par rapport à ces différentes fonctions rhétoriques.

3.7 Quelle suite pour ce travail ?

Afin que nos annotations puissent être incluses dans le corpus du projet Democrat, qui requiert des textes qu'on peut modifier et diffuser librement, nous serons sans doute amené à changer notre propre corpus, et donc à abandonner l'étude des textes de format IMRaD. Néanmoins, les outils et la méthodologie que nous avons élaborés, ainsi que notre caractérisation des chaînes de paragraphe au chapitre 6, ne dépendent pas d'un corpus, et pourront servir à étudier les chaînes de référence d'autres types de texte.

Annexe A

Description et extraits du corpus

1 Description des textes du corpus

Nous donnons ici la liste des textes, avec le résumé afin que les exemples inclus dans ce travail soit plus parlants.

Texte T0 : Dagot L., Borteyrou X., Grégoire C., Vallée B. (2014). Le rôle modérateur des compétences politiques sur le burnout. *Revue internationale de psychologie sociale*, 27 (2), www.cairn.info/revue-internationale-de-psychologie-sociale-2014-2-page-5.htm, consulté le 14 janvier 2016. Cet article traite de deux objectifs : premièrement l'adaptation d'une échelle française de mesure des compétences politiques en milieu organisationnel, sur la base de l'échelle de Ferris *et al.* (2005), et deuxièmement la mise en évidence du rôle modérateur des compétences politiques sur le burnout. L'adaptation de l'échelle a été réalisée auprès d'une population de 170 salariés français. L'étude du rôle modérateur des compétences politiques sur le niveau de burnout a mobilisé une population de 496 salariés français. La version française de l'échelle de compétences politiques correspond à la structure de l'échelle initiale, et l'analyse confirmatoire offre des indicateurs d'ajustement très satisfaisants. Le rôle modérateur des compétences politiques est en partie vérifié. Plus précisément, les résultats indiquent l'importance de bien distinguer les quatre dimensions des compétences politiques. Les résultats de ces deux études encouragent la poursuite des recherches sur les compétences politiques en contexte francophone. Cela nécessite de poursuivre le processus de validation de l'échelle en français.

Texte T1 : Herrscher E., Lheureux J., Goude G., Dabernat H., Duranthon F. (2016). Les pratiques de subsistance de la population Néolithique final de la grotte I des Treilles (commune de Saint-Jean-et-Saint-Paul, Aveyron). *Préhistoires Méditerranéennes*, 4, http://pm.revues.org/783, consulté le 29 janvier 2016. La Grotte I des Treilles (commune de Saint-Jean-et-Saint-Paul, Aveyron) se situe sur l'extrémité sud-occidentale du Causse du Larzac. C'est une grotte sépul-

crale à caractère collectif et des datations radiocarbone l'ont attribuée au Néolithique final (3030-2890 cal B.C, phase moyenne du Groupe des Treilles). Cette recherche a pour objectif (1) d'appréhender le type d'écosystème exploité par la population de la Grotte I des Treilles pour subvenir à ses besoins nutritionnels et (2) d'étudier les relations entre les pratiques alimentaires et des données anthropométriques. Pour cela, une analyse isotopique a été menée sur 42 individus adultes (pour un total de 86 individus adultes mis au jour) ainsi que sur 14 ossements de faune découverts sur plusieurs sites archéologiques contemporains régionaux. Les résultats isotopiques montrent une consommation locale des protéines par la population de la Grotte I des Treilles. La contribution des protéines animales apparait importante relativement aux protéines végétales. Nos résultats suggèreraient également la contribution de viande de jeunes animaux non encore sevrés dans l'alimentation. La quantité de protéines animales consommée par cette population est très variable entre les individus du groupe. Toutefois, cette diversité de comportement alimentaire n'est pas corrélée aux données anthropométriques (longueurs des os longs et robustesse). Les choix alimentaires n'apparaissent pas liés aux paramètres biologiques testés et ne sont donc probablement pas dictés par le sexe des individus.

Texte T2 : Houy T., Attal Y., Melamed Y. (2016). La propension à l'interdisciplinarité des étudiants en situation d'innovation. *Revue internationale de pédagogie de l'enseignement supérieur*, 30 (2), http://ripes.revues.org/825 , consulté le 31 janvier 2016. L'article vise à qualifier le comportement des étudiants lorsqu'ils sont confrontés à l'interdisciplinarité en situation d'innovation. Nous montrons que les étudiants disposent d'une appétence pour les travaux de groupe réalisés avec d'autres élèves du supérieur spécialisés dans d'autres disciplines. Cette propension à l'interdisciplinarité diffère selon le type d'étudiant et les établissements. Les résultats avancés dans cet article proviennent du traitement des données issues de la plateforme Studyka. Cette plateforme en ligne permet à des étudiants de se mettre en relation, quels que soient leur spécialité et leur établissement pour travailler ensemble et relever un challenge autour de l'innovation.

Texte T3 : Hulin G., Broes F. et Fechner K. (2012). Caractérisation de phénomènes anthropiques par la mesure de paramètres magnétiques sur surface décapée : Premiers résultats sur le projet Canal Seine-Nord Europe. *ArcheoSciences*, 36, http://archeosciences.revues.org/3744 , consulté le 26 mars 2016. La démarche suivie sur le projet Canal Seine-Nord Europe se veut différente de l'utilisation standard qui peut être faite de la géophysique. La volonté est d'intégrer cet outil au plus près de l'étude archéologique comme peuvent l'être la pédologie ou la cartographie des phosphates par exemple. Grâce à une interaction forte avec les archéologues et géo-archéologues du projet, de nouvelles problématiques ont été définies et mises en oeuvre. La plus intéressante est certainement l'emploi de la géophysique pour l'étude des surfaces décapées. Par cette approche, il s'agit non pas de détecter les structures mais d'apporter à l'archéologue une information nouvelle, complémentaire des données de fouilles. Par la mesure d'un ou de plusieurs paramètres magnétiques, des phénomènes anthropiques peuvent être mis en évidence et apporter des éléments pour la compréhension du site. Les différents cas présentés illustrent la diversité des problématiques pouvant faire intervenir la géophysique. Celles-ci sont nombreuses et l'apport scientifique particulièrement important. Tout en gardant un regard critique, l'étude sur surface décapée offre un très fort potentiel qui doit encore être développé. La comparaison des résultats géophysiques avec la cartographie des phosphates constitue un des axes majeurs de cette démarche.

Texte T4 : Dernat S., Siméone A. (2014). Représentations socio-professionnelles et choix de la spécialisation : le cas de la filière vétérinaire rurale. *Revue internationale de pédagogie de l'enseignement supérieur*, 30 (2) http://ripes.revues.org/832 , consulté le 26 mars 2016. Cet article éclaire comment une représentation sociale spécifique, la représentation socio-professionnelle, permet d'expliquer le choix de la spécialisation d'étudiants dans l'enseignement supérieur, en dépassant la seule prise en compte de critères sociodémographiques. Pour illustrer notre propos, nous avons investigué la problématique du choix de la filière « rurale » dans l'enseignement vétérinaire français qui fait face à une désaffection marquée depuis de nombreuses années. Les résultats dégagés à la suite d'une enquête quantitative ($n = 533$) croisant plusieurs méthodologies suggèrent que le cursus d'apprentissage et, en particulier, les stages pourraient jouer un rôle important dans le passage d'une représentation sociale à une représentation professionnelle. Ces résultats remettent en cause la simple prise en compte des critères de recrutement des étudiants et mettent en avant l'importance de l'ingénierie pédagogique dans la construction représentationnelle et les choix de la spécialisation.

2 Exemples d'annotation

À titre d'exemple, nous mettons ici les deux derniers paragraphes du texte T2, d'abord avec l'annotation du chapitre 5 (« Une première étude exploratoire : annotation d'une sélection de référents saillants »), puis avec l'annotation du chapitre 6 (« Une deuxième étude exploratoire : annotation systématique des chaînes de paragraphe »). Nous mettons à la disposition du jury l'ensemble des annotations, au format utilisé par Analec.

Le format des annotations ci-dessous est le suivant : index, catégorie, fonction, et éventuellement expansion(s). Les codes sont ceux décrits dans le schéma d'annotation (chapitre 4, page 80). Par exemple « [cet article]$_{\text{i:m,c}}$ » signifie que l'index est i, qu'il s'agit d'un SN démonstratif complément circonstantiel, sans expansion.

Exemple d'annotation pour le chapitre 5 :

> Au moins deux pistes pourraient être ouvertes pour confirmer les résultats liminaires contenus dans [cet article]$_{\text{i:m,c}}$. Une première piste consisterait à conduire une enquête quantitative, sur une base déclarative, de manière à interroger [les étudiants]$_{\text{j:d,v}}$ sur les raisons qui les poussent vers [l'interdisciplinarité]$_{\text{k:d,v}}$. Cette démarche présenterait les inconvénients habituels des traitements statistiques réalisés sur données déclaratives. Elle compléterait néanmoins [le présent article]$_{\text{i:d,v,a}}$ et pourrait ouvrir des pistes nouvelles permettant d'expliquer [l'intérêt [des étudiants]$_{\text{j:d,n}}$ pour les autres disciplines]$_{\text{l:d,v,nn}}$. Une deuxième piste consisterait à produire une analyse qualitative en conduisant une série d'entretiens auprès d'une population représentative d'[étudiants]$_{\text{j:t,n}}$. L'analyse de ces échanges [nous]$_{\text{m:s,v}}$ permettrait de vérifier en particulier l'exhaustivité des hypothèses formulées dans [cet article]$_{\text{i:m,c}}$ pour justifier [la propension à [l'interdisciplinarité]$_{\text{k:d,n}}$ [des étudiants]$_{\text{j:d,n}}$]$_{\text{l:d,v,nn}}$.
>
> Ces prolongements permettraient de mieux comprendre [le comportement [des étudiants]$_{\text{j:d,n}}$ face à [l'interdisciplinarité]$_{\text{k:d,c}}$]$_{\text{n:d,v,nn}}$, notamment en situation d'[innovation]$_{\text{o:t,n}}$. En fonction des résultats de ces travaux, les enseignants pourraient envisager les formes pédagogiques les plus adaptées pour enseigner [l'innovation]$_{\text{o:d,v}}$ et [l'entrepreneuriat]$_{\text{p:d,v}}$. L'utilité d'une telle recherche serait d'autant plus importante que les méthodes

pédagogiques pour enseigner [l'entrepreneuriat]$_{p:d,v}$ et [l'innovation]$_{o:d,v}$ restent peu évaluées et qu'il ne semble pas exister de modèle dominant, même si la pratique de la pédagogie active est la plus employée (Fayolle & Verzat, 2009).

Exemple d'annotation pour le chapitre 6 :

Au moins deux pistes pourraient être ouvertes pour confirmer les résultats liminaires contenus dans [cet article]$_{i:m,c}$. Une première piste consisterait à conduire une enquête quantitative, sur une base déclarative, de manière à interroger [les étudiants]$_{j:d,v}$ sur les raisons qui les poussent vers l'interdisciplinarité. [Cette démarche]$_{k:m,s}$ présenterait les inconvénients habituels des traitements statistiques réalisés sur données déclaratives. [Elle]$_{k:s,s}$ compléterait néanmoins [le présent article]$_{i:d,v,a}$ et [Ø]$_{k:z,s}$ pourrait ouvrir des pistes nouvelles permettant d'expliquer l'intérêt [des étudiants]$_{j:d,n}$ pour les autres disciplines. Une deuxième piste consisterait à produire une analyse qualitative en conduisant une série d'entretiens auprès d'une population représentative d'[étudiants]$_{j:t,n}$. L'analyse de ces échanges nous permettrait de vérifier en particulier l'exhaustivité des hypothèses formulées dans [cet article]$_{i:m,c}$ pour justifier la propension à l'interdisciplinarité [des étudiants]$_{j:d,n}$.

Ces prolongements permettraient de mieux comprendre le comportement des étudiants face à l'interdisciplinarité, notamment en situation [d'innovation]$_{l:t,n}$. En fonction des résultats de ces travaux, les enseignants pourraient envisager les formes pédagogiques les plus adaptées pour enseigner [l'innovation]$_{l:d,v}$ et l'entrepreneuriat. L'utilité d'une telle recherche serait d'autant plus importante que les méthodes pédagogiques pour enseigner l'entrepreneuriat et [l'innovation]$_{l:d,v}$ restent peu évaluées et qu'il ne semble pas exister de modèle dominant, même si la pratique de la pédagogie active est la plus employée (Fayolle & Verzat, 2009).

Annexe B

Liste des métadonnées

1 Implémentation dans le texte

Nous incluons les métadonnées directement dans le fichier du texte. Chacune est sur une ligne, qui commence par le symbole `#`. Cela permet de les repérer facilement, et de les éliminer des décomptes (nombre de tokens, de paragraphes, etc.).

Voici un exemple d'insertion des métadonnées :

```
#journal: ArchoSciences

#volume: 36

#year: 2012

#title: Caractérisation de phénomènes anthropiques (...)

#authors: Hulin Guillaume, Broes Frédéric et Fechner Kai

#url: http://archeosciences.revues.org/3744

#accessdate: 2016-03-26

#textid:t0003

...

#part-type:introduction
```

```
#part-heading:level=1

Introduction

Les études géophysiques (...)

(...)

**********************************************************************

#part-type:methodology

#part-heading:level=1

Méthodes

#part-heading:level=2

Mesure directe de la susceptibilité magnétique

De manière générale, (...)
```

2 Liste des métadonnées

Certaines métadonnées sont des indications bibliographiques :

- `journal` : nom de la revue,
- `volume` : numéro de la publication,
- `year` : année,
- `title` : titre de l'article,
- `authors` : auteurs,
- `country` : pays,
- `base` : base bibliographique (par exemple `revues.org`),
- `url` : url,
- `accessdate` : date d'accès,
- `keyword` : mots-clés (une métadonnée par mot-clé),
- `abstract` : résumé.

D'autres concernent la gestion des textes et des référents annotés :

- `textid` : numéro du texte,
- `referent` : nom du référent (une métadonnée par mot-clé),
- `additionnaltoken` : permet d'ajouter un token spécial (par exemple s'il contient des espaces, des tirets, des caractères spéciaux).

D'autres encore permettent le découpage en partie, et le balisage des inter-titres :

- `part-heading` : le paragraphe suivant est un inter-titre, avec indication du niveau hiérarchique (`level=1`),
- `part-type` : le type de partie (par exemple `introduction`),
- une ligne de `*` : séparateur de partie.

Enfin, certaines permettent d'indiquer la présence de tableaux ou d'illustrations (qui ont été ignoré lors de l'annotation) :

- `graphic` : légende de l'illustration,
- `graphic-notes` : notes de l'illustration,
- `table` : légende du tableau,
- `table-notes` : notes du tableau.

Annexe C

Guide d'utilisation de l'interface

1 Aperçus...

1.1 Fichiers d'essai

Pour tester les fonctionnalités de l'interface, il suffit d'utiliser les données d'essai, en cliquant sur les boutons suivants de la page de démarrage :

- `aesop` (une fable d'Ésope),
- `aesop (empty)` (la même fable mais sans les annotations),
- `papyrus` (un papyrus de Zénon du IIIe siècle avant notre ère),
- `papyrus (empty)` (le même papyrus mais sans les annotations) ;

puis sur le bouton *Parse the data* (si on maintient `ctrl` enfoncé en cliquant sur le bouton, on parse toute de suite le texte).

1.2 Aperçu du mode opératoire de l'annotation

Le mode opératoire est assez simple :

- Pour **créer un maillon**, il suffit de sélectionner un ou plusieurs mots, puis de cliquer sur le bouton correspondant à ce que l'on veut faire :
  - créer une nouvelle chaîne avec le maillon,
  - attacher le nouveau maillon à une chaîne existante.
- La **gestion des chaînes** est automatique : dès qu'un ensemble de maillons contient trois éléments ou plus, l'ensemble devient une « vraie » chaîne, obtient une couleur et un bouton d'accès rapide. Tant que l'ensemble n'a qu'un ou deux maillons (c'est-à-dire tant qu'il est un singleton ou une paire), il a une couleur par défaut et apparaît dans une liste à part. Mais tout cela est automatique.

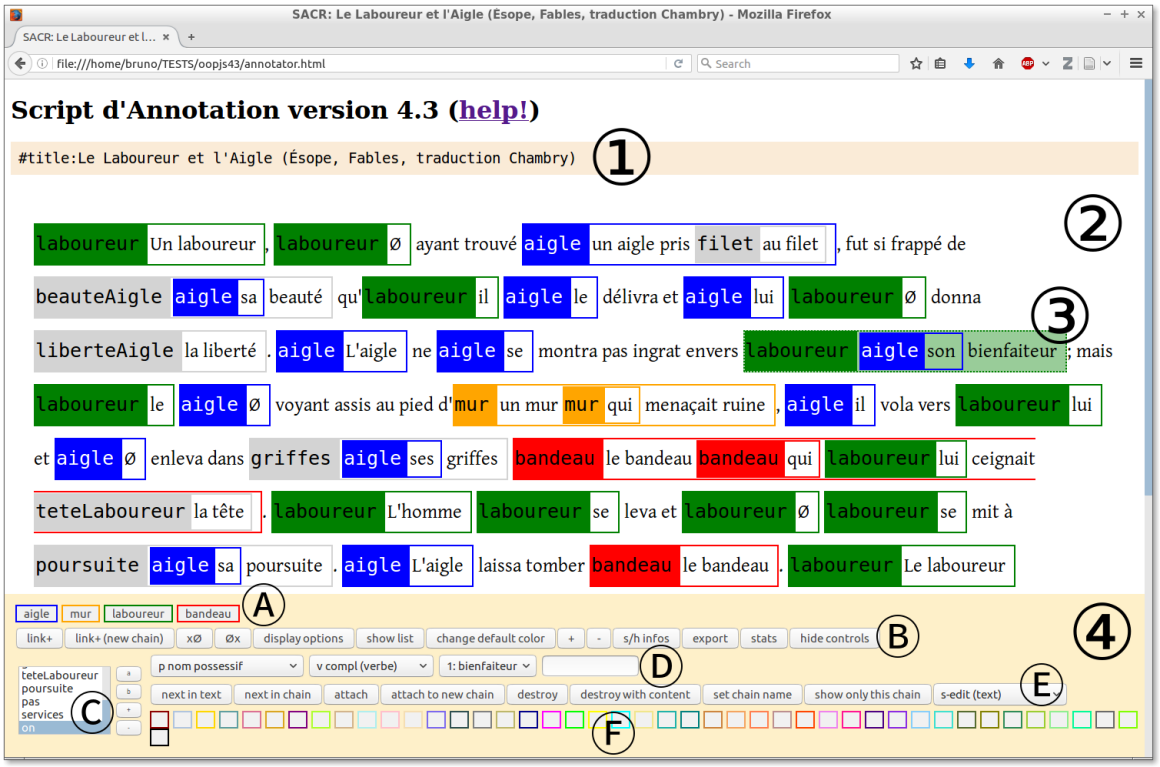

Fig. 10: Interface du script d'annotation

Attention : Dans la suite, nous réserverons le terme « chaîne » pour les ensembles de trois maillons ou plus, celui de « singleton/paire » pour les ensembles d'un ou deux maillons, et celui d'« ensemble » pour parler indifféremment des chaînes ou des singletons/paires.

1.3 Aperçu de l'interface

La figure 10 illustre l'interface et ses principales parties :

- **1 :** Les métadonnées sont affichées en police « monotype » et sur fond jaune.
- **2 :** Texte annoté. Le nom de la chaîne (*i.e.* du référent) peut être affiché ou non. Dans ce dernier cas, seul les cadres apparaissent.
- **3 :** Exemple d'un maillon sélectionné (il a une couleur de fond semi-transparente).
- **4 :** Le panneau de commandes :
  - **A :** Boutons d'accès rapides aux chaînes (qui ont plus de trois maillons). Il suffit de cliquer sur l'un de ces boutons colorés pour (créer au besoin et) attacher le maillon à la chaîne.
  - **B :** Commandes générales de création de maillons, de chaînes, d'options d'affichage, d'exportation des annotations, etc.
  - **C :** Liste des singletons/paires. Il est possible d'augmenter/réduire le nombre d'éléments affichés dans la liste (boutons `+` et `-`), mais aussi de trier les chaînes par ordre alphabétique (bouton *a*) ou par ordre d'apparition (bouton *b*), ce qui est pratique si le texte est long et s'il y a beaucoup de singletons/paires, et qu'on veut vérifier si un référent est déjà dans la liste.

- **D :** Liste des propriétés pour le maillon actuellement sélectionné (ce panneau est caché si aucun maillon n'est sélectionné).
- **E :** Commandes pour le maillon sélectionné ou pour la chaîne à laquelle appartient le maillon sélectionné (ce panneau est caché si aucun maillon n'est sélectionné).
- **F :** Boutons qui permettent facilement de changer la couleur de la chaîne à laquelle appartient le maillon sélectionné (lorsqu'un ensemble contient au moins trois maillons, il automatiquement promu au rang de « chaîne » et une couleur lui est automatiquement attribuée. Il peut donc être utile de lui attribuer une autre couleur par la suite).

1.4 Singletons/paires, chaînes et chaînes spéciales

Il y a trois types d'ensembles :

- Les **singletons/paires**, qui ont un ou deux maillons, ne sont pas (encore) des chaînes de référence. Ils le deviendront automatiquement dès qu'ils auront trois maillons ou plus. Ils ont tous une même couleur par défaut (qu'il est possible de modifier). Ces ensembles n'ont pas de bouton coloré spécial, mais sont listés dans la liste en bas à gauche de l'interface.
- Les **chaînes** ont trois maillons ou plus. Chacune a sa couleur et son bouton coloré d'accès rapide. Si une chaîne perd certains de ses maillons (soit parce qu'on les a supprimés, soit parce qu'on les a attachés à d'autres chaînes), et qu'elle a moins de trois maillons, elle est déchue de son rang de « chaîne », perd automatiquement sa couleur et son bouton d'accès rapide, et se retrouve listée dans la liste en bas à gauche, avec les autres singletons/paires.
- Les **chaînes spéciales** commencent par un caractère de soulignement (_). Ces ensembles ont toujours une couleur et un bouton, même s'ils ont moins de trois maillons. En fait, ce mécanisme a été créé pour annoter d'autres éléments que des chaînes de référence (par exemple des marqueurs discursifs, ou des verbes). Ces éléments peuvent recevoir des propriétés différentes.

Lorsqu'un ensemble (quel que soit son type) perd la totalité de ses maillons, il est automatiquement supprimé.

2 Format des fichiers de données

2.1 Encodage et fins de ligne

L'encodage à utiliser est UTF-8, avec ou sans BOM. Pour convertir un texte vers UTF-8, rien de plus simple : plutôt que charger le fichier directement, il suffit de le copier dans la zone de texte prévue à cet effet. Le texte récupéré par Javascript est automatiquement converti en UTF-8.

Les fins de ligne peuvent être de type Windows ou Unix. Lors de l'exportation, le programme détecte (en lisant la valeur de `userAgent`) si l'utilisateur utilise Windows. Dans ce cas, il offre le choix entre les fins de ligne Windows (par défaut) ou Unix. Sinon, les fins de ligne sont celles d'Unix.

```
# définition d'une nouvelle propriété, "categorie", attachée
# à toutes les chaînes qui ne commencent pas par "_":
PROP:name=categorie,target="*regular"
# la ligne suivante signifie un "blanc": l'utilisateur peut donc
# choisir de laisser la propriété non remplie
$$$
nom propre
SN indéfini
SN défini
...

# autre propriété: "fonction":
PROP:name=fonction,target="*regular"
$$$
sujet
objet
...

# autre propriété, de type "label", attachée à toutes les chaînes (pas
# d'option "target").  Cette propriété n'a pas de valeur (c'est une
# textbox qui s'affiche)
PROP:name=label,type=label

# autre propriété, de type "text", attachée uniquement à la "chaîne
# spéciale" nommée _marqueur.  Il s'agit du type du marqueur, qu'on peut
# entrer librement dans une textbox.
PROP:name=typeMarqueur,showname=true,type=text,target="^_marqueur"
```

FIG. 11: Exemple de schéma.

2.2 Les propriétés

2.2.1 Le format

Le premier document à définir est celui qui contient la liste des propriétés et leurs valeurs possibles.

Les propriétés sont les paramètres annotés, comme la catégorie grammaticale ou la fonction grammaticale (voir figure 11).

Le format est donc le suivant (attention : pas d'espace autorisé, sauf entre des guillemets) : `PROP:opt1=value1,opt2="value2 with space",....`

Options et valeurs possibles :

- `name` (obligatoire) : nom de la propriété (libre) ;
- `type` (optionel) :

  - `normal` : une liste déroulante (il faut mettre des valeurs à la suite). C'est la valeur par défaut ;
  - `head` : tête syntaxique. Aucune valeur à fournir : la liste est calculée automatiquement à partir des mots contenus dans le maillon. Cette propriété permet de calculer automatiquement la distance intermaillonnaire (tête à tête) et les coefficients de stabilité ;
  - `label` : identifiant du maillon. Aucune valeur à fournir : c'est une *textbox* qui s'affiche. La valeur n'est jamais vide : par défaut, il s'agit du nombre de milliseconde depuis l'Epoch, sinon n'importe quel texte entré par l'utilisateur. S'il n'entre rien, c'est le nombre de millisecondes qui est inséré automatiquement. Il est conseiller de ne pas modifier la valeur par défaut. Attention : pas de contrôle d'unicité.
  - `text` : Aucune valeur à fournir : c'est une *textbox* qui s'affiche. L'utilisateur peut entrer ce qu'il veut ;
  - `ref` : comme `text`, mais l'utilisateur est censé entrer la valeur `label` d'un autre maillon. Attention : pas de contrôle de validité ;
- `showname` (optionel) : faut-il afficher le nom de la propriété ?
  - `true` : oui,
  - `false` : non (défaut) ;
- `newline` (optionel) : faut-il commencer une nouvelle ligne après la propriété (simple question d'affichage) ?
  - `true` : oui,
  - `false` : non (défaut) ;
- `textboxsize` (optionel) : taille de la *textbox* (7 par défaut) ;
- `target` (optionel) : à quelle chaîne la propriété est-elle associée ? La valeur peut être :
  - un raccourci :
    - `*all` : toutes les chaînes (équivalent de `""`, l'expression régulière qui réussit sur toutes les chaînes de caractères). C'est le cas par défaut,
    - `*regular` : uniquement les chaînes qui ne commencent *pas* par `_` (équivalent de `^[^_]`),
    - `*special` : uniquement les chaînes qui *commencent* par `_` (équivalent de `^_`) ;
  - une expression régulière testée sur le nom de la chaîne. Par exemple `^_marqueur$` va s'appliquer sur la chaîne nommée `_marqueur`, `^[^_]` sur toutes les chaînes non-spéciales, `^_` sur toutes les chaînes spéciales, etc.

Le mécanisme `label/ref` permet de gérer des relations entre chaînes (fusion, séparation, etc.). Il existe un moyen facile de remplir ces propriétés automatiquement :

- Quand on **déselectionne** un maillon en maintenant `ctrl` enfoncé, la valeur de la propriété `label` (si la propriété existe pour le maillon) est stockée.
- Quand on **sélectionne** un maillon en maintenant `shift` enfoncé, la valeur `label` précédemment stockée est copiée dans le champ `ref` du maillon (si la propriété existe pour le maillon).

Ce qui veut dire que pour copier le label d'un maillon sélectionné dans le maillon qu'on veut sélectionner ensuite, il faut tenir `ctrl` et `shift` enfoncés en même temps.

2.2.2 Chargement du fichier

Cliquer sur le bouton *Properties: Browse...* , ou bien copier (ou taper directement) le texte dans la zone de texte correspondante.

2.2.3 Changement des propriétés après l'annotation

Il est possible de changer les propriétés après l'annotation. Dans ce cas, le programme détecte que des propriétés sont en trop, ou manquantes, ou ont des valeurs incorrectes. Il affiche des messages d'erreur et laisse les propriétés nouvelles ou incorrectes avec des valeurs par défaut. Pour ne pas afficher ces messages, il faut décocher la case *show property warnings* .

2.3 Le texte

2.3.1 Format

Le deuxième document à créer est le fichier qui contient le texte à annoter (voir figure 12).

Il s'agit d'un fichier en texte brut, qui peut contenir les informations suivantes :

- Les lignes commençant par `#` sont des métadonnées, qui ne sont pas destinées à être annotées mais qui donnent des renseignements (source, provenance, titre, date, etc.). Ces informations apparaissent dans l'interface en police monotype et sur fond jaune. Deux métadonnées servent à afficher le titre de la fenêtre (`#title` et `#textid`).
- Les lignes de plusieurs étoiles (`*`) ou de plusieurs `#` sont des séparateurs (des différentes parties) et apparaissent comme des lignes.
- Les lignes `#additionnaltoken:new token` permettent de définir des tokens particulier, notamment s'ils contiennent des espaces ou des caractères spéciaux.
- Les lignes `#part-heading:level=1` indique que le prochain paragraphe est un titre de section, dont le niveau est indiqué par `level` (on peut aller de 1 à 3). L'affichage du paragraphe qui suit est donc plus grand, plus large, plus gras, etc.
- Les lignes de format `#COLOR:green:ReferentName` sont des lignes utilisées pour sauver la couleur pour la chaîne donnée. Elles sont écrites et lues automatiquement par le programme.
- La ligne `#DEFAULTCOLOR:red` a le même but.
- Les autres lignes sont des lignes de texte. Les paragraphes doivent être séparés par des lignes blanches. Il peut y avoir des sauts de ligne à l'intérieur d'un paragraphe (mais pas de ligne blanche).

Les annotations sont codées selon un format « XML déguisé » : `{ReferentName:prop1=val1,prop2="text with space" TEXTE}`

Par exemple :

```
{chat:categorie="SN défini",fonction="sujet" Le petit chat de
{voisine:categorie="SN défini",fonction="cdn" la voisine}}
```

```
#title:Le Laboureur et l'Aigle (Ésope, Fables, traduction Chambry)

{laboureur:categorie="i nom indéfini",fonction="s sujet",head="1:
laboureur",expansion="" Un laboureur}, {laboureur:categorie="z pronom
zéro",fonction="s sujet",head="0: ∅",expansion="" ∅} ayant trouvé
(...)

**********************************************************************

Il faut rendre (...)

#COLOR:green=laboureur
#COLOR:blue=aigle
#COLOR:orange=mur
#COLOR:red=bandeau
#DEFAULTCOLOR:gray
```

Fig. 12: Exemple de fichier texte.

Les guillemets ne sont pas obligatoires si les caractères correspondent à des caractères latins (`[a-zA-Z]`) ou des chiffres.

Les annotations sont donc récursives et on peut imbriquer autant d'annotations que l'on veut (mais l'affichage ne sera optimal que pour quatre niveaux : à partir du cinquième, les cadres se superposeront. Il est possible d'augmenter le nombre, mais l'espace inter-ligne devient alors plus important).

Les annotations sont donc des annotations « embarquées » lisibles par un humain. Ce qui permet de corriger facilement des fautes d'orthographe, d'ajouter du texte ou d'en supprimer. Dans le pire des cas, cela permet de récupérer facilement les données et de les « réparer » en cas de problème avec l'interface. De plus, l'utilisateur peut lire les données sans même avoir recours à l'interface. Enfin, cela permet d'utiliser d'autres scripts, en Perl ou Python par exemple, pour des calculs statistiques plus avancés, des conversions, des extractions, etc.

2.3.2 Chargement du fichier

Cliquer sur le bouton *Text: Browse...* , ou bien copier (ou taper directement) le texte dans la zone de texte correspondante.

Fig. 13: Importation des données.

3 Utilisation de l'interface

3.1 Importation des fichiers

Il faut commencer par écrire, copier ou charger les propriétés et le texte qu'on veut annoter, comme indiqué précédemment (voir figure 13).

3.2 Création des maillons

Pour créer un maillon, il faut sélectionner des mots (plus exactement des *tokens*) en cliquant dessus : le mot est alors souligné. Pour créer un maillon d'un seul mot, on sélectionne un seul mot. Pour créer un maillon de plusieurs mots, on sélectionne le premier mot et le dernier mot du maillon. On ne peut pas sélectionner plus de deux mots.

Il y a ensuite plusieurs possibilités :

- Soit on veut créer le maillon et l'attacher à un singleton qui n'existe pas encore : on clique sur *link+ (new chain)* . Le programme demande alors le nom du singleton (un nom basé sur le contenu du maillon est proposé par défaut).
- Soit on veut créer le maillon et l'attacher à un singleton/paire qui existe déjà. Il faut sélectionner le nom dans la liste en bas à gauche et cliquer sur *link+* . Si on vient d'ajouter le troisième maillon, l'ensemble est promu au rang de « chaîne » : il obtient automati-

quement une couleur (qu'on peut changer) et un bouton coloré en haut du panneau de contrôle.
- Soit on veut créer le maillon et l'attacher à une chaîne, il suffit de cliquer sur le bouton coloré correspondant en haut du panneau de contrôle.

3.3 Modification des maillons et des chaînes

Pour sélectionner un maillon, il faut cliquer dessus (sur le nom du référent, ou dans l'espace à l'intérieur du cadre mais pas sur un mot, car sinon on sélectionne le mot). Le panneau de contrôle affiche alors les éléments qui permettent de modifier le maillon ou la chaîne associée :

- Les **propriétés** : il suffit de sélectionner l'élément dans la liste, ou bien de taper du texte à l'intérieur d'une textbox. La propriété se met automatiquement à jour. La figure 14 montre le changement de la propriété `head` .
- Bouton *next in text* : sélectionne le prochain maillon dans le texte. Un clic avec `ctrl` enfoncé permet de sélectionner le maillon *précédent*. Un clic avec `shift` enfoncé permet de sélectionner le maillon suivant (ou précédent, si `ctrl` et `shift` sont enfoncés en même temps) qui est *visible* (ce qui est pratique pour parcourir tous les maillons trouvées après une recherche).
- Bouton *next in chain* : sélectionne le prochain maillon dans la chaîne. Même combinaison de touches que le bouton précédent.
- Bouton *attach* : attache le maillon à un autre singleton/paire : il faut l'avoir préalablement sélectionné dans la liste en bas à gauche. Pour attacher le maillon à une chaîne qui dispose d'un bouton coloré en haut du panneau, il suffit de cliquer sur ce bouton.
- Bouton *attach to a new chain* : attache le maillon à un singleton/paire qui n'existe pas et est créé pour l'occasion (le nom est demandé, un nom est proposé par défaut).
- Bouton *destroy* : enlève le maillon (mais pas le texte !). Pour corriger les bornes d'un maillon, voir ci-dessous.
- Bouton *destroy with content* : détruit le maillon et le texte à l'intérieur. Cela n'est possible que pour les maillons qui ne contiennent qu'un « Ø » (c'est la seule façon d'enlever un « Ø » créer par erreur).
- Bouton *set chain name* : permet de donner un nouveau nom à l'ensemble (singleton, paire ou chaîne) du maillon sélectionné.
- Bouton *show only this chain* : affiche les métadonnées de tous les maillons de l'ensemble, et cache celles de tous les autres.
- Case *edit mode* : active le mode d'édition (voir plus bas).
- Les **boutons colorés** au bas du panneau de contrôle permettent de changer la couleur de la chaîne du maillon sélectionné.

Attention à ne pas confondre sélection d'un mot (mot souligné) et sélection d'un maillon (maillon avec un fond coloré à moitié transparent). La sélection d'un mot permet de créer des maillons et d'insérer des symboles « Ø ». La sélection des maillons permet de les annoter. Les deux types sélections sont mutuellement exclusives.

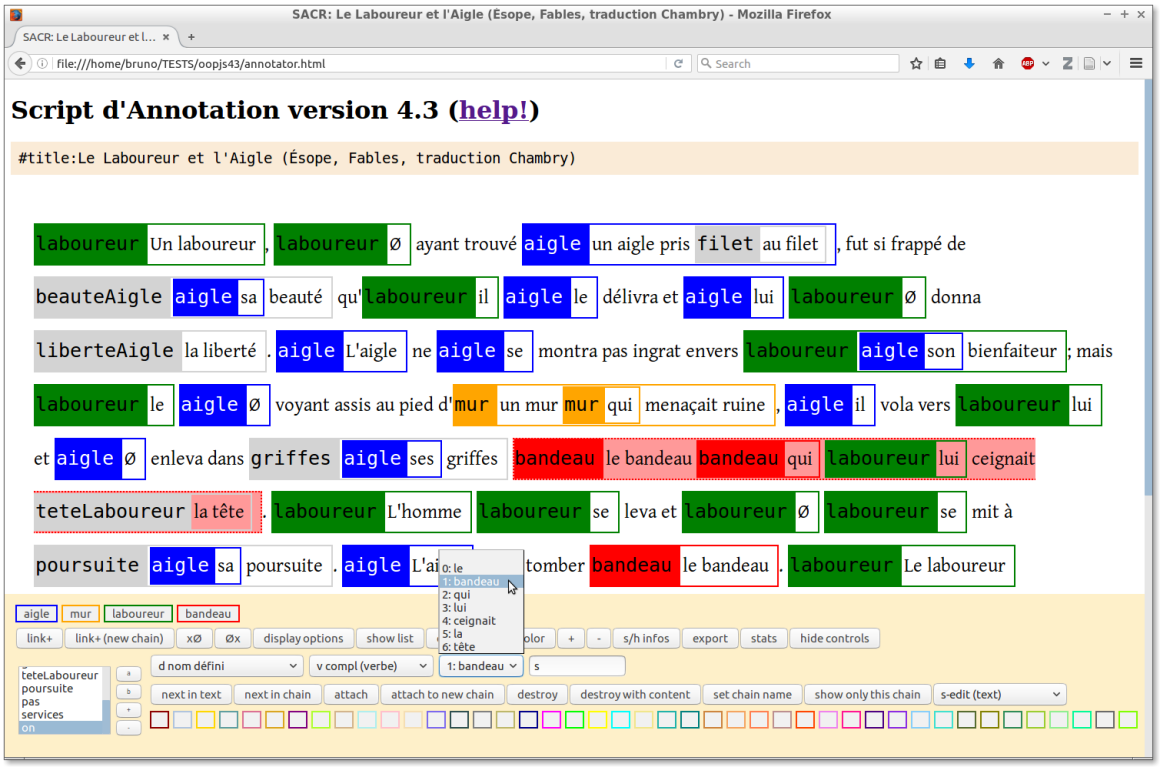

FIG. 14: Interface du script d'annotation, avec sélection d'une valeur de la propriété *head*..

Pour corriger les bornes du maillon, il faut sélectionner le ou les mots (comme pour créer un nouveau maillon), puis sélectionner le maillon à corriger tout en maintenant la touche `ctrl` enfoncée.

3.3.1 Mode édition

Pour annoter rapidement des maillons, c'est-à-dire définir leurs propriétés, on peut activer des modes spéciaux, dits *edit modes* . Pour les utiliser, il vaut mieux faire commencer les valeurs de propriétés par une lettre unique, par exemple `d SN défini` , `i SN indéfini` , etc. Lorsque le focus est sur une liste déroulante, enfoncer la touche correspondante à la propriété voulue (par exemple `d`) va sélectionner cette propriété (par exemple `d SN défini`).

Il y a deux modes d'édition. Le premier est activé en sélection *edit mode* dans la liste déroulante des modes. Il permet d'annoter rapidement toutes les propriétés, maillon après maillon. Le principe est le suivant : on sélectionne le premier maillon, puis on donne le focus à la première propriété. On enfonce alors la touche correspondante à la propriété (`d` dans notre exemple), et la propriété voulue (`d SN défini`) se sélectionne automatiquement. Le focus va alors directement à la propriété suivante, et on recommence (`s` pour `s sujet` par exemple). Lorsqu'on a parcouru toutes les propriétés, on arrive sur le bouton *next in text* , avec lequel on peut sélectionner le maillon suivant (en appuyant sur la barre d'espace). Le focus revient alors automatiquement à la première propriété, et on recommence.

Le deuxième mode d'édition est appelé *super edit* . Le principe est similaire, mais on annote cette fois-ci une propriété pour tous les maillons, puis la propriété suivante pour tous les maillons, etc. L'annotation devient extrêmement rapide :

- un maillon est mis en relief, c'est-à-dire en couleur (et le texte défile automatiquement pour mettre le maillon vers le centre de l'écran au besoin),
- il suffit d'appuyer sur la touche correspondante à la valeur choisie (`d` dans notre exemple),
- automatiquement le maillon suivant est mis en relief (et se présente au centre de l'écran si besoin), et on recommence.

Il n'y a donc aucun clic de souris ni aucune touche qui est enfoncée inutilement : une frappe de touche correspond à une annotation ; on ne peut pas faire plus économique...

On peut choisir quatre modes *super edit* . On peut ainsi parcourir tous les maillons d'une chaîne (pour annoter chaîne par chaîne) ou de toutes les chaînes, et on peut choisir de parcourir tous les maillons ou seulement ceux qui sont visibles (ce qui permet, par exemple, de n'annoter que ceux qui n'ont pas encore reçu d'annotations).

3.3.2 Relations entre maillons

Le mécanisme `label/ref` présenté plus haut permet de gérer des relations entre maillons ou entres chaînes. Par exemple, il est possible de spécifier que tel maillon (« l'église » dans les phrases : « Nous arrivâmes dans un village. L'église était en hauteur. »), qui initie une nouvelle chaîne, a une relation d'anaphore associative avec un autre maillon (« un village » dans l'exemple précédent).

Outre les anaphores associatives, ce système permet aussi de gérer les « partie de », ou bien les fusions de chaînes, divisions, etc.

Comme dit plus haut, il existe un moyen facile de remplir ces propriétés automatiquement :

- Quand on **déselectionne** un maillon en maintenant `ctrl` enfoncé, la valeur de la propriété `label` (si la propriété existe pour le maillon) est stockée.
- Quand on **sélectionne** un maillon en maintenant `shift` enfoncé, la valeur `label` précédemment stockée est copiée dans le champ `ref` du maillon (si la propriété existe pour le maillon).

Ce qui veut dire que pour copier le label d'un maillon sélectionné dans le maillon qu'on veut sélectionner ensuite, il faut tenir `ctrl` et `shift` enfoncés en même temps.

Cette relation n'est pas représentée graphiquement. (Bien qu'il soit à l'étude de pouvoir afficher seulement les maillons d'une telle « chaîne liée », au sens informatique du terme.)

3.4 Mode de visualisation

3.4.1 Un peu de terminologie

Les chaînes peuvent être visualisées de différente façon :

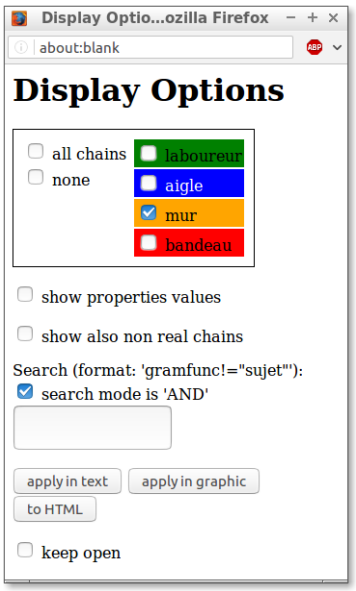

FIG. 15: La boîte de dialogue des options d'affichage.

- dans le texte (bouton *apply to text* de la fenêtre *display options*),
- dans un « graphique » (bouton *apply to graphic*),
- dans un nouveau fichier HTML (bouton *to HTML*).

Qu'est-ce qu'un maillon « affiché » ou « caché » ?

- Dans le texte :
  - maillon affiché : le maillon a le nom du référent devant lui, sur fond coloré,
  - maillon chaché : pas de nom de référent, seulement un cadre de couleur qui rappelle sa présence ;
- Dans le « graphique » :
  - maillon affiché : le maillon est souligné de couleur,
  - maillon caché : rien du tout n'est affiché ;
- Dans un nouveau fichier HTML :
  - maillon affiché : le maillon a le nom du référent devant lui,
  - maillon chaché :
    - si `ctrl` est enfoncé au moment où le fichier HTML est créé (*i.e.* au moment où on clique sur le bouton *to HTML* dans la fenêtre *display options*) : pas de cadre,
    - sinon : un cadre de couleur autour du maillon.

Les figures 16 et 17 illustrent ces modes d'affichage.

Noter qu'un maillon qui a juste un cadre autour de lui (*i.e.* caché) est tout de même sélectionnable : il faut juste cliquer dans l'espace blanc à l'intérieur du cadre (mais pas sur un mot). Cela permet, par exemple, d'annoter sans être déranger par le nom des référents (qui parfois gêne la bonne lecture du texte).

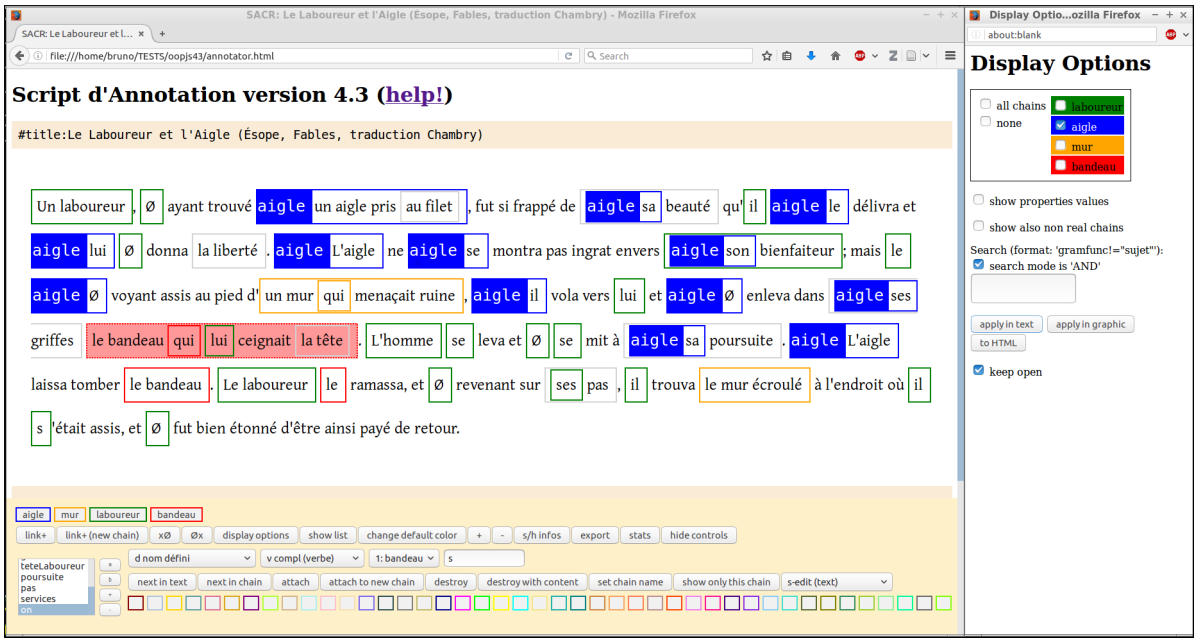

FIG. 16 : Affichage de la chaîne *aigle*, et seulement elle. Noter que le maillon caché *le bandeau qui lui ceignait la tête* est sélectionné.

3.4.2 La boîte de dialogue *display options*

La boîte de dialogue *display options* (figure 15) est accessible en cliquant sur le bouton *display options* .

Sélection des chaînes

Le premier cadre permet de sélectionner les chaînes à afficher. On peut toutes les afficher (*all chains*), n'en afficher aucune (*none*), ce qui permet, par exemple, de lire le texte sans être déranger par le nom des référents, ou bien en afficher que certaines (pour faire apparaître la liste, il faut décocher *all chains* et *none*) ; c'est cette dernière option qu'illustre la figure 16.

Propriétés

L'option *show properties values* affiche à côté du nom du référent l'ensemble des propriétés. Cela n'est utile que pour le débogage.

Les singletons/paires

L'option *show also non real chains* permet d'afficher (ou de masquer) les singletons/paires. Ils sont traités en bloc : on ne peut pas en afficher un séparément.

Recherche

Le bloc suivant permet d'affiner l'affichage. Il permet de sélectionner seulement certains maillons en fonction du contenu des propriétés (voir la figure 17).

- L'option *search mode is "AND"* active le mode `AND` (c'est-à-dire que tous les critères spécifiés dans la recherche doivent être remplis) si elle est cochée, ou au contraire le mode `OR` (un seul critère a besoin d'être rempli) si elle est décochée.

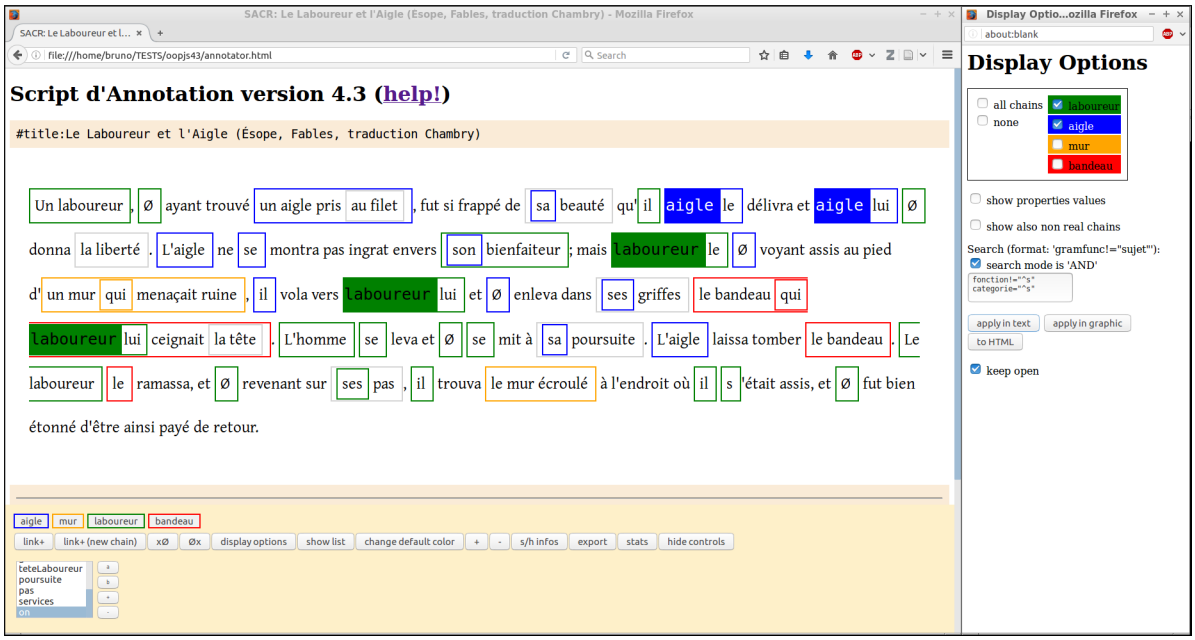

FIG. 17 : Affichage des maillons qui sont : (1) des pronoms personnels, (2) *non* sujet, (3) des chaînes *laboureur* ou *aigle* (mais pas des autres).

- La zone de texte permet de spécifier les critères de recherche. Le format est le suivant : `NOM_PROPRIETE="REGEX"` ou `NOM_PROPRIETE!="REGEX"` (noter le point d'exclamation), où
  - `NOM_PROPRIETE` est le nom d'une propriété (tel que définie dans le fichier des propriétés). Pour ce souvenir des noms des propriétés, il suffit de maintenir enfoncés `ctrl+shift` et de cliquer sur *export* dans le panneau de contrôle de la fenêtre principale : cela affiche la liste des propriétés et leurs valeurs,
  - `=` ou `!=` est l'opérateur (chercher ce qui correspond au critère pour le premier, ce qui *ne* correspond *pas* au critère pour le second).
  - `REGEX` est une expression régulière qui est testée sur la valeur de la propriété (ex. : `sujet`) Il faut noter que `""` est automatiquement remplacé par `^$` , l'expression régulière qui trouve les propriétés vides, ce qui est utile à la fin de l'annotation pour vérifier qu'on n'a pas oublié d'annoter l'un ou l'autre maillon.

Il est bien sûr possible de mettre plusieurs de ces critères (qui seront recherchés en mode `AND` ou `OR`), en les séparant par des espaces (ou des sauts de ligne, ce qui permet d'en mettre un par ligne et d'améliorer la visibilité).

Voici un exemple, qui recherche tous les indéfinis *non* sujets en mode `AND` , ou bien tous les indéfinis mais aussi tous les maillons *non* sujest en mode `OR` (dans ce cas, le mode `AND` a plus de sens !) : `categorie="indéfini" fonction!="sujet"` .

Il n'est pas possible, pour l'heure, de faire des recherches avec priorité, ni de mêler le mode `AND` et le mode `OR` . Il est néanmoins possible de cocher l'option `AND` , et faire du `OR` à l'intérieur de chaque propriété grâce à la puissance des expressions régulières, par exemple : `categorie="indéfini" fonction="(sujet|objet)"` .

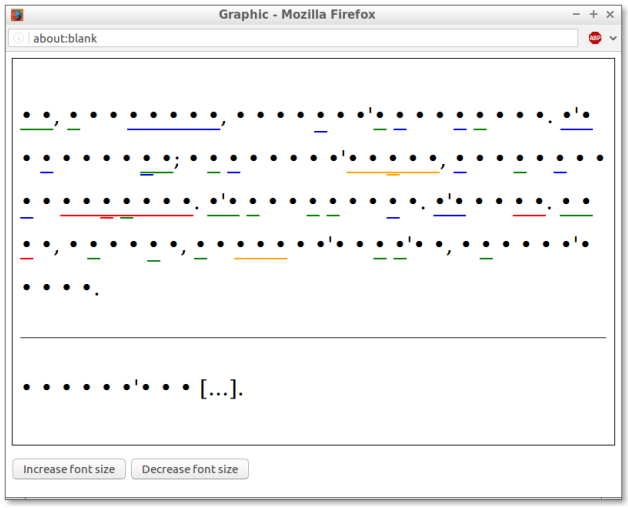

FIG. 18: Représentation « grahique » du texte.

Les boutons

Les boutons permettent d'appliquer les options d'affichage dans le texte, dans le graphique ou dans un nouveau fichier HTML.

Le principe de l'affichage « graphique » (voir figure 18) est le suivant : Tous les mots sont remplacés par un gros point (« • »), et les différents maillons sont soulignés (sur quatre niveaux selon les paramètres actuels du script). Cet affichage permet de voir la distribution des maillons en couleur, sur tout le texte. On peut voir par exemple si une chaîne apparaît plus au début ou à la fin, si les maillons ont beaucoup de mots, etc. Si tous les sujets, par exemple, ont la même couleur, cela signifie que tous les maillons appartiennent à la même chaîne.

Il s'agit juste d'un affichage, on ne peut pas cliquer sur les maillons.

Par ailleurs, le bouton *to HTML* supporte les touches suivantes :

- `ctrl` : masquer totalement les maillons non affichés, ne pas dessiner de cadre coloré autour,
- `shift` : afficher tous les commentaires (lignes qui commencent par `#`). Par défaut, seule la ligne `#title` et les lignes de séparateurs de parties sont exportées en HTML.

Pour laisser la boîte ouverte

L'option *keep open*, enfin, permet de laisser la fenêtre ouverte. Sinon, elle se ferme dès qu'on appuie sur l'un des boutons.

3.4.3 Liste des chaînes et des maillons

Avec le bouton *show list*, on affiche une fenêtre (voir figure 19) qui liste toutes les chaînes et tous les maillons, dans l'ordre d'apparition. On peut cliquer sur un maillon pour le sélectionner : il est alors aussi sélectionné dans la fenêtre principale, et inversement. En maintenant

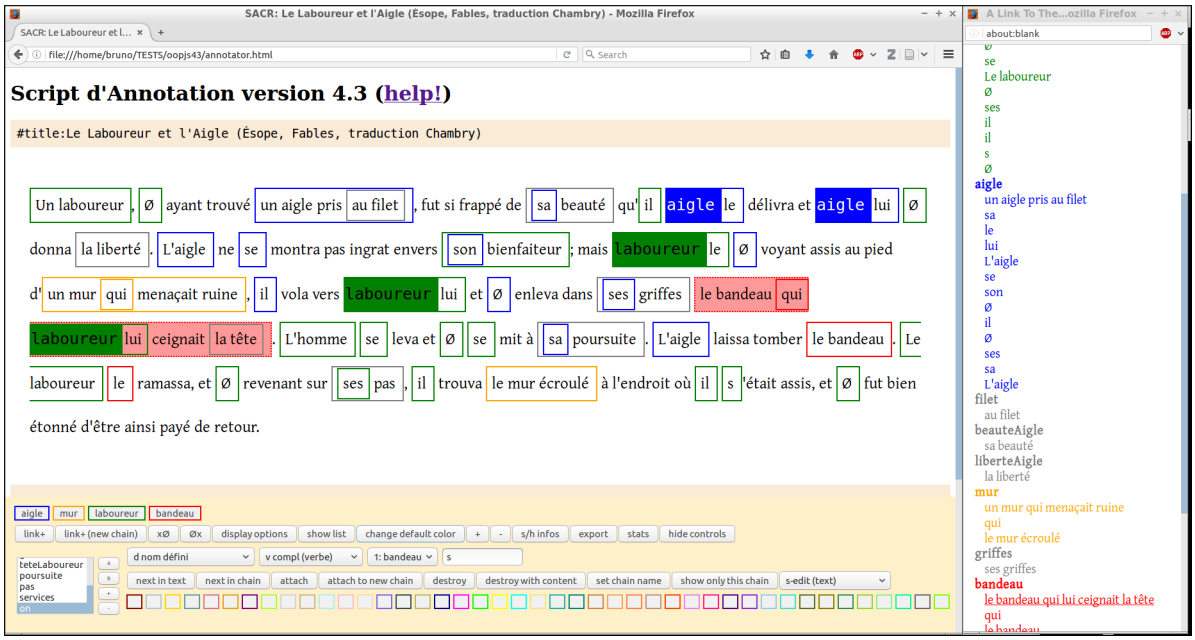

FIG. 19: La boîte de dialogue qui liste l'ensemble des maillons.

`ctrl` enfoncé quand on clique sur le maillon, la fenêtre principale se déroule jusqu'au maillon. Cela est utile quand on a un long texte.

La liste se met à jour automatiquement quand on ajoute, supprime ou renomme des maillons ou des chaînes. Cependant, il est plus prudent, pour l'heure, de fermer la fenêtre avant de se livrer à de grandes modifications...

3.5 Autres commandes

Il reste à voir les boutons suivants :

- Boutons *xØ* et *Øx* : insèrent le symble d'ensemble vide pour signifier un « pronom zéro », avant ou après le mot sélectionné. Il agit comme n'importe quel mot : on peut le sélectionner et créer un maillon avec. Pour supprimer ce symbole, il faut créer un maillon le contenant et supprimer le maillon avec le bouton *destroy with content* .
- Bouton *export* : exporte le texte et les annotations. **Attention : C'est le seul moyen de « sauvegarder » les annotations : il faut les exporter !** On peut exporter dans le même fichier, mais le programme propose par défaut un nom accompagné de la date et de l'heure, ce qui permet de sauvegarder plusieurs versions du même fichier, et de revenir en arrière en cas d'erreur. Sous Windows, le script demande s'il doit utiliser les fins de ligne spéciales Windows, ou rester avec celles d'Unix. On peut se servir des touches de contrôle :
  - `ctrl` : exporte non le texte, mais le schéma (= les propriétés),
  - `shift` : exporte dans une boîte de dialogue (attention : pour les textes longs, la boîte de dialogue n'affiche pas la fin du texte : c'est une limitation des navigateurs),
  - `ctrl+shift` : exporte le schéma dans une boîte de dialogue.
- Bouton *set default color* : change la couleur des singletons/paires.

- Bouton *s/h infos* : affiche ou masque les métadonnées.
- Bouton *stats* : affiche quelques statistiques de bases (nombre de mots, de chaînes, de maillons, etc.).
- Boutons `+` et `-` : change la taille de la police.
- Boutons *hide controls* : masque le panneau de contrôle. Pour le réafficher, cliquer sur le bouton tout au bas de la page.

Attention : Avant de fermer la page web ou de quitter le navigateur, il faut veiller à sauvegarder le texte annoté au moyen du bouton *export* .

4 Compatibilité

Le script a été conçu et testé sur Firefox (à partir de la version 43). Il devrait tourner sur Chromium/Chrome. Il ne tourne pas sous Internet Explorer.

Annexe D

Implémentations

1 Liste des scripts

Voici la liste des scripts que nous avons écrits (sauf mention contraire, ils sont écrits en Perl) :

- l'interface d'annotation des chaînes de référence, en HTML, CSS et Javascript,
- les scripts d'analyses des annotations (statistiques et fréquences des propriétés),
- les scripts qui nous ont permis de télécharger les plans d'articles sur `revues.org` pour la recherche d'articles au format IMRaD,
- le script qui calcule les patrons les plus fréquents,
- les scripts de conversions de et vers le format Glozz, ce qui permet de convertir de et vers Analec,
- divers scripts d'ajout, suppression et modification des annotations.

2 Implémentations

Au vu de la longueur des codes écrits, nous ne les avons pas inclus dans l'édition de ce travail. Nous les tenons cependant à la disposition du jury, sous forme électronique.

Bibliographie

Abréviation :

- GMF : Riegel, Pellat et Rioul, 2014.

Ariel M. (1990). *Accessing noun-phrase antecedents*. London ; New York : Routledge.

Asher N. (1993). *Reference to abstract objects in discourse*. Dordrecht, Boston : Kluwer Academic.

—— (2000). Events, Facts, Propositions, and Evolutive Anaphora. *In* Varzi A., Higginbotham J. et Pianesi F. *Speaking of Events*. Oxford University Press.

Bazerman C. (1988). *Shaping written knowledge : the genre and activity of the experimental article in science*. Madison : University of Wisconsin Press.

Benayoun J.-M. (2003). Sujet Ø, pacte référentiel et thème. *In* Merle J.-M. (éd.) *Le sujet : actes, augmentés de quelques articles, du Colloque Le sujet*. Gap, Paris : Ophrys.

Bessonnat D. (1988). Le découpage en paragraphes et ses fonctions. *Pratiques*, 57, 81–105.

Boure R. (1998). Produire une revue scientifique. Le cas de Sciences de la Société. *In* Renzetti F. (éd.) *Stratégies informationnelles et valorisation de la recherche scientifique publique*. Paris : ADBS.

Brett P. (1994). A genre analysis of the results section of sociology articles. *English for Specific Purposes*, 13 (1), 47–59.

Capin D. (2014). Chaînes de référence dans les textes médiévaux non-narratifs : les Year Books ou l'élaboration d'une écriture juridique. *Langages* (195), 61–78.

Carnie A. (2013). *Syntax : a generative introduction*. 3e édition. Hoboken, New Jersey, USA : Wiley-Blackwell.

Charolles M. (1988). Les plans d'organisation textuelle : périodes, chaînes, portées et séquences. *Pratiques*, 57, 3–13.

—— (2002). *La référence et les expressions référentielles en français*. Paris : Ophrys.

—— (2007). Comment évaluer les effets des relatives en qui sur les chaînes de référence ?. *In* Fournier N., Charolles M., Fuchs C. et Lefeuvre F. *Parcours de la phrase : mélanges offerts à Pierre Le Goffic*. Paris : Ophrys.

Charolles M. et Schnedecker C. (1993). Coréférence et identité : le problème des référents évolutifs. *Langages*, 27 (112), 106–126.

Chastain C. (1975). Reference and context. *In* Gunderson K. (éd.) *Language, mind, and knowledge*. Minneapolis : University of Minnesota Press.

Choi-Jonin I. et Delhay C. (1998). *Introduction à la méthodologie en linguistique : application au français contemporain*. Strasbourg : Presses Universitaires de Strasbourg.

Condette M.-H., Marin R. et Merlo A. (2012). La structure argumentale des noms déverbaux : du corpus au lexique et du lexique au corpus. *SHS Web of Conferences*, 845–858.

Corblin F. (1985a). Les chaînes de référence : analyse linguistique et traitement automatique. *Intellectica*, 1 (1), 123–143.

—— (1985b). Remarques sur la notion d'anaphore. *Revue québécoise de linguistique*, 15 (1).

—— (1987). *Indéfini, défini et démonstratif : constructions linguistiques de la référence*. Genève : Droz.

—— (1995). *Les formes de reprise dans le discours : anaphores et chaînes de référence*. Rennes : Presses universitaires de Rennes.

—— (1997). Les indéfinis : variables et quantificateurs. *Langue française*, 116 (1), 8–32.

—— (2002). *Représentation du discours et sémantique formelle : introduction et applications au français*. Paris : Presses universitaires de France.

—— (2013). *Cours de sémantique : introduction*. Paris : A. Colin. 1 volumes.

Croft W. et Cruse D. A. (2004). *Cognitive linguistics*. Cambridge, New York : Cambridge University Press.

Cruse D. A. (1986). *Lexical semantics*. Cambridge, New York : Cambridge University Press.

—— (2000). *Meaning in language : an introduction to semantics and pragmatics*. Oxford, New York : Oxford University Press.

Davidson D. (2004 [1966]). The Logical Form Of Action Sentences. *In* Davis S. et Gillon B. S. (éds) *Semantics : a reader*. New York : Oxford University Press.

—— (2006 [1969]). The individuation of events. *In* Lepore E. et Ludwig K. (éds) *The essential Davidson*. Oxford : Clarendon Press : Oxford University Press.

Delaveau A. (2001). *Syntaxe : la phrase et la subordination*. Paris : A. Colin.

Fløttum K. (2006a). The typical research article—does it exist ?. *In* Suomela-Salmi E. et Dervin F. (éds) *Cross-linguistic and cross-cultural perspectives on academic discourse*. Amsterdam, Philadelphia : John Benjamins.

—— (2006b). Les « personnes » dans le discours scientifique : le cas du pronom ON. *XVIe Congrès des Romanistes Scandinaves*. Department of language and Culture, Roskilde University. http://rudar.ruc.dk:8080/bitstream/1800/8139/1/Artikel33.pdf, consulté le 2016-06-01.

Fløttum K., Jonasson K. et Norén C. (2007). *On : pronom à facettes*. Bruxelles [Paris] : De Boeck-Duculot. 1 volumes.

Fort K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. thèse de doctorat de l'Université Paris 13.

Frege G. (1971b [1891]). Fonction et Concept. *In* Frege G. *Écrits logiques et philosophiques*. Paris : Éditions du Seuil. Traduit par Imbert Claude. Originellement publié en 1891.

—— (1971c [1892]). Concept et Objet. *In* Frege G. *Écrits logiques et philosophiques*. Paris : Éditions du Seuil. Traduit par Imbert Claude. Originellement publié en 1891.

Giry-Schneider J. (1987). *Les Prédicats nominaux en français : les phrases simples à verbe support*. Genève : Droz.

Glikman J., Guillot-Barbance C. et Obry V. (2014). Les chaînes de référence dans un corpus de textes narratifs médiévaux : traits généraux et facteurs de variation. *Langages* (195), 43–60.

Gross G. (2006). Prédicats, anaphores et classes d'objets. *In* Riegel M., Schnedecker C., Swiggers P. et Tamba I. *Aux carrefours du sens : hommages offerts à Georges Kleiber pour son 60e anniversaire*. Louvain, Paris : Peeters.

—— (2012). *Manuel d'analyse linguistique : approche sémantico-syntaxique du lexique*. Villeneuve-d'Ascq : Presses universitaires du Septentrion.

Gross G. et Vivès R. (2001). La description en termes de classes d'objets et l'enseignement des langues. *Langue française*, 38–51.

Grossmann F. (2010). L'Auteur scientifique. *Revue d'anthropologie des connaissances*, 4 (3), 410–426.

—— (2012). Pourquoi et comment cela change ? Standardisation et variation dans le champ des discours scientifiques. *Pratiques. Linguistique, littérature, didactique*, 153-154, 141–160.

Habert B. (2005). *Instruments et ressources électroniques pour le français*. Gap, Paris : Ophrys.

Habert B., Nazarenko A. et Salem A. (1997). *Les linguistiques de corpus*. Paris : A. Colin.

Heim I. (1982). *The semantics of definite and indefinite noun phrases*. Schoubye/Glick.

Heslot J. (1983). Récit et commentaire dans un article scientifique. *Documentation et recherche en linguistique allemande contemporaine*, 29, 133–54.

Ho-Dac L.-M. (2005). Deux modes de segmentation textuelle : univers de discours et chaînes de référence. *Verbum* (3), 231–248.

Huang Y. (2000). *Anaphora : a cross-linguistic approach*. Oxford ; New York : Oxford University Press.

—— (2014). *Pragmatics*. 2e édition. Oxford : Oxford University Press.

Huyghe R. (2012). Noms d'objets et noms d'événements : quelles frontières linguistiques ?. *Scolia* (26), 81–103.

—— (2014). La sémantique des noms d'action : quelques repères. *Cahiers de lexicologie*, 105, 181–201.

Jacques M.-P. (2005). Structure matérielle et contenu sémantique du texte écrit. *Corela. Cognition, représentation, langage*, 3 (2).

—— (2013). Structure textuelle de l'article scientifique. *In* Tutin A. et Grossmann F. (éds) *L'écrit scientifique : du lexique au discours*. Rennes : Presses universitaires de Rennes.

Karttunen L. (1976). Discourse Referents. *In* McCawley J. D. (éd.) *Notes from the linguistic underground*. New York : Academic Press.

Kleiber G. (1981). *Problèmes de référence : descriptions définies et noms propres*. Metz : Université de Metz, Centre d'analyse syntaxique.

—— (1997). Sens, référence et existence : que faire de l'extra-linguistique ?. *Langages*, 9–37.

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

—— (2014). Anaphores et coréférences : analyse assistée par ordinateur. *In* Fossard M. et Béguelin M.-J. (éds) *Nouvelles perspectives sur l'anaphore : points de vue linguistique, psycholinguistique et acquisitionnel*. Bern : Peter Lang.

—— (2016). Conception d'un outil de visualisation et d'exploration de chaînes de coréférences. *JADT 2016 : 13ème Journées internationales d'Analyse statistique des Données Textuelles*.

Landragin F., Poibeau Th. et Victorri B. (2012). ANALEC : a new tool for the dynamic annotation of textual data. *International Conference on Language Resources and Evaluation (LREC 2012)*. https://halshs.archives-ouvertes.fr/halshs-00698971/

Landragin F. et Tanguy N. (2014). Référence et coréférence du pronom indéfini on. *Langages* (195), 99–115.

Langacker R. W. (2008). *Cognitive grammar : a basic introduction*. Oxford, New York : Oxford University Press.

Le Goffic P. (1993). *Grammaire de la phrase française*. Paris : Hachette supérieur.

Leeman D. (2002). *La phrase complexe : les subordinations*. Bruxelles : De Boeck-Duculot.

Lemaréchal A. (1997). *Zéro(s)*. Paris : Presses Universitaires de France.

Loffler-Laurian A.-M. (1980). L'expression du locuteur dans les discours scientifiques. *Je*, *Nous*, et *On* dans quelques textes de chimie et de physique. *Revue de Linguistique Romane Lyon* (173), 135–157.

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 l'Université de Strasbourg.

Longo L. et Todirascu A. (2013). Une étude de corpus pour la détection automatique de thèmes. *Revue électronique Texte et corpus*, 4, 143–155.

—— (2014). Vers une typologie des chaînes de référence dans des textes administratifs et juridiques. *Langages* (195), 79–98.

—— (2015). La saillance référentielle pour la détection des thèmes. *In* Boisseau M. et Hamm A. (éds) *Saillance , Volume 2 : La saillance en langue et en discours*. Besançon : Presses universitaires de Franche-Comté.

Lundquist L., Minel J.-L. et Couto J. (2012). La navigation discursive. L'anaphore réomptive et mouvement discursif. *L'analyse du discours dans la société*, 365–389.

Milard B. (2007). La mise en forme des publications scientifiques : entre routines, contraintes et organisation de l'expérience collective. *In* Gaudez F. et Peuchlestrade G. (éds) *Sociologie des arts, sociologie des sciences : actes du colloque international de Toulouse, 2004*. Paris : L'Harmattan.

Milner J.-C. (1978). Le système du réfléchi en latin. *Langages*, 12 (50), 73–86.

—— (1982). *Ordres et raisons de langue*. Paris : Éditions du Seuil.

Moreau M.-L. (1976). *C'est : étude de syntaxe transformationnelle*. Mons, Belgique : Editions Universitaires de Mons.

de Mulder W. (1998). Du sens des démonstratifs à la construction d'univers. *Langue française*, 120 (1), 21–32.

Muzerelle J., Lefeuvre A., Schang E., Antoine J.-Y., Pelletier A., Maurel D., Eshkol I. et Villaneau J. (2014). ANCOR Centre, a large free spoken French coreference corpus : description of the resource and reliability measures. *LREC'2014, 9th Language Resources and Evaluation Conference*.

Mélanie-Becquet F. et Landragin F. (2014). Linguistique outillée pourl'étude des chaînes de référence : questions méthodologiques et solutions techniques. *Langages* (195), 117–137.

Müller-Gjesdal A. (2013). The Influence of Genre Constraints on Author Representation in Medical Research Articles. The French Indefinite Pronoun On in IMRAD Research Articles. *Discours. Revue de linguistique, psycholinguistique et informatique*, 12.

Neveu F. (2004). *Dictionnaire des sciences du langage*. Paris : A. Colin.

Perret M. (2000). Quelques remarques sur l'anaphore nominale aux XIVe et XVe siècles. *L'Information Grammaticale*, 87 (1), 17–23.

Pontille D. (2003). Formats d'écriture et mondes scientifiques. Le Cas de la Sociologie. *Questions de communication*, 3, 55–67.

—— (2007). Matérialité des écrits scientifiques et travail de frontières : le cas du format IMRAD. *Sciences et frontières*, 229–253.

Poudat C. (2006). *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*. Thèse de l'Université d'Orléans.

Quine W. V. (2013 [1960]). *Word and object*. Cambridge : MIT Press. Originellement publié en 1960.

Regent O. (1992). Pratiques de communication en médecine : contextes anglais et français. *Langages*, 26 (105), 66–75.

Riegel M., Pellat J.-C. et Rioul R. (2014). *Grammaire méthodique du français*. 5e édition. Paris : Presses Universitaires de France.

Rinck F. (2006). *L'article de recherche en Sciences du langage et en Lettres : figure de l'auteur et identité disciplinaire du genre*. Thèse de l'Université Stendhal de Grenoble.

—— (2010). L'analyse linguistique des enjeux de connaissance dans le discours scientifique : un état des lieux. *Revue d'anthropologie des connaissances*, 4 (3), 427–450.

Régent O. (1980). Approche comparative des discours de spécialité pour l'entraînement à l'anglais écrit. *Mélanges Pédagogiques du CRAPEL*.

Schnedecker C. (1997). *Nom propre et chaînes de référence*. Metz : Université de Metz.

—— (2005). Les chaînes de référence dans les portraits journalistiques : éléments de description. *Travaux de linguistique*, 51 (2), 85–133.

—— (2014). Chaînes de référence et variations selon le genre. *Langages*, 195 (3), 23–42.

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

Schnedecker C. et Longo L. (2012). Impact des genres sur la composition des chaînes de référence : le cas des faits divers. *3ième Congrès Mondial de LInguistique Française*, 1957–1972.

Seuren P. A. M. (1998). *Western linguistics : an historical introduction*. Oxford : Blackwell Publishers.

Swales J. M. (1990). *Genre analysis : English in academic and research settings*. Cambridge, New York : Cambridge University Press.

—— (2004). *Research genres : explorations and applications*. Cambridge, New York : Cambridge University Press.

Tellier C. (1995). *Éléments de syntaxe du français : méthodes d'analyse en grammaire générative*. Montréal : Presses de l'Univ. de Montréal.

Tutin A. (2002). A corpus-based study of pronominal anaphoric expressions in French. *Proceedings of DAARC 2002 (Discourse Anaphora and Anaphora Resolution), Lisbon, 18-20 september 2002*. http://agnes.tutin.u-grenoble3.fr//Publis/DAARC2002.pdf, consulté le 2016-06-03.

Tutin A. et Grossmann F. (éds) (2013). *L'écrit scientifique : du lexique au discours*. Rennes : Presses universitaires de Rennes.

Victorri B. (2011). Analec : logiciel d'annotation et d'analyse de corpus écrits téléchargeable sur : http://www.lattice.cnrs.fr/-Analec- .

Widlöcher A. et Mathet Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. *In* Nazarenko A., Poibeau Th. et Audibert L. *Actes de la conférence Traitement Automatique des Langues Naturelles*. ATALA.

—— (2012). The glozz platform : A corpus annotation and mining tool. *In Proceedings of the 2012 ACM Symposium on Document Engineering*. New York : Association for Computing Machinery.

Williams I. A. (1999). Results Sections of Medical Research Articles : Analysis of Rhetorical Categories for Pedagogical Purposes. *English for Specific Purposes*, 18 (4), 347–366.

Table des matières