Linguistique outillée pour l'étude des chaînes de référence : questions méthodologiques et solutions techniques

Frédérique Mélanie-Becquet, Frédéric Landragin

▶ **To cite this version:**

Frédérique Mélanie-Becquet, Frédéric Landragin. Linguistique outillée pour l'étude des chaînes de référence : questions méthodologiques et solutions techniques. Langages, 2014, 195, pp.117-137. halshs-01069462

HAL Id: halshs-01069462
https://shs.hal.science/halshs-01069462

Submitted on 29 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Linguistique outillée pour l'étude des chaînes de référence : questions méthodologiques et solutions techniques

– DRAFT –

Frédérique Mélanie-Becquet, Frédéric Landragin
Lattice, CNRS/ENS/Université de Paris 3 Sorbonne Nouvelle

1. Introduction

Les outils et instruments informatiques sont désormais indispensables à la linguistique de corpus (Habert *et al.*, 1997). Que l'annotation soit entièrement manuelle ou combine une part manuelle et une part automatique, il existe une quantité d'outils capables d'enrichir un texte d'annotations de toutes sortes : morphologiques, syntaxiques, sémantiques, pragmatiques. Dès que l'on s'intéresse aux phénomènes de référence et de coréférence, plusieurs questions méthodologiques se posent. Tout d'abord, les phénomènes de référence font appel au contexte, contexte que le lecteur se construit au fur et à mesure de la lecture, et qui peut entraîner des biais interprétatifs. Pour que les annotations ne soient pas trop subjectives, un manuel d'annotation strict et directif s'avère nécessaire. Il faut cependant que le schéma d'annotation tienne compte des ambiguïtés et flous possibles, et autorise une certaine souplesse dans l'affectation des valeurs. Ensuite, l'outil utilisé doit permettre une construction simple et efficace du monde des référents : annotations des relations entre chacun des référents en présence, gestion de chaînes de référence (qui sont susceptibles de couvrir le texte entier). Il est indispensable d'avoir un outil adapté tant pour la saisie des chaînes, leur visualisation, la gestion et l'interrogation des données annotées, que pour la réalisation de calculs statistiques.

Dans cet article, nous faisons une synthèse des questions posées et des difficultés rencontrées lors de l'étude des chaînes de référence menée dans le cadre du projet MC4 « Modélisation Contrastive et Computationnelle des Chaînes de Coréférence » (projet CNRS à l'origine de ce volume, cf. section 4 de l'article de présentation). Le projet a permis d'initier le travail d'annotation de chaînes de référence sur un corpus de textes écrits. Dans la section 2, nous présentons rapidement les besoins d'annotation du projet : définition du corpus, des phénomènes linguistiques à étudier, et par conséquent du schéma d'annotation. Dans la section 3, nous nous interrogeons sur la réalisation pratique de ce schéma d'annotation : quel format lui donner, comment le gérer. Dans la section 4, nous présentons quelques solutions techniques adaptées, présentes dans l'outil ANALEC – ANAlyse de L'ECrit, cf. (Landragin *et al.*, 2012) – et notamment dans la version 1.4, retravaillée pour le projet MC4. Dans la dernière section, nous abordons les limites de cette version de l'outil, et de ce fait les consolidations ou améliorations de certaines fonctionnalités qui seraient opportunes et envisageables.

2. Le projet MC4 et ses besoins d'annotation

2.1. Le corpus

Le corpus écrit du projet MC4 comprend 9 textes, soit environ 20 000 mots – une taille ni trop minimale ni trop importante, qui permet un traitement manuel et qualitatif des données (cf. tableau 1). Ce corpus a été constitué par les membres participants du projet MC4 en fonction de leur connaissance préalable des textes voire d'études antérieures sur d'autres phénomènes. Le projet MC4 a pour but d'annoter les phénomènes référentiels, à savoir un ensemble défini (mais non exhaustif) d'indices présents dans le texte, quelle que soit leur nature, du nom propre au pronom (cf. la section 2.2 de l'article de présentation de ce volume, avec une liste de formes et l'échelle d'accessibilité associée). Chacun de ces indices est nommé « maillon » et entre dans la constitution d'une « chaîne de référence » (CR). Ainsi dans le premier paragraphe du second chapitre de *La mère Sauvage* de Maupassant, où apparaît le personnage éponyme, sont annotés 10 maillons, répartis sur 3 CR correspondant à 3 référents humains : le fils, la mère et un référent collectif indéfini. Ci-dessous, nous soulignons chacune des formes constituant un maillon. Le pronom *qui* constitue une forme à lui tout seul, et appartient de plus à une autre forme, ce qui explique le double soulignement. Par ailleurs, le sujet non exprimé du participe *laissant* donne lieu lui aussi au repérage d'un maillon :

> *Lorsque la guerre fut déclarée, le fils Sauvage, qui avait alors trente-trois ans, s'engagea, Ø laissant la mère seule au logis. On ne la plaignait pas trop, la vieille, parce qu'elle avait de l'argent, on le savait.*

Titre, Auteur	Siècle	Type	Nombre de mots de l'extrait annoté	Nombre de maillons annotés	Nombre de maillons pour 100 mots
Gracial d'Adgar	12e	Vers	2641	631	24
Quatre Livres des Rois	12e	Prose	2211	470	21
La vie de Saint Thomas Becket	12e	Vers	2067	485	23
Li Estoires de Chiaus qui conquisent Coustantinoble, Robert de Clari	12e – 13e	Prose	1639	245	15
Queste del saint Graal	13e	Prose	2224	450	20
Les quinze Joyes de mariage, Premiere joye	14e – 15e	Prose	2457	630	26
Les Bijoux, Maupassant	20e	Prose	2448	418	17
La Mère sauvage, Maupassant	20e	Prose	2450	503	21
L'occupation des sols, Echenoz	20e	Prose	1787	234	13
Total :			19924	4066	

Tableau 1. Liste des œuvres qui constituent le corpus écrit MC4.

L'ensemble des textes réunis n'est pas homogène puisque constitué de textes en vers ou en prose, d'époques différentes, de longueur variable, correspondant ou non à l'ensemble de

l'œuvre (à savoir : 6 récits du *Gracial d'Adgar*, le premier livre des *Quatre Livres des Rois*, etc.). Cette hétérogénéité s'explique par une volonté de réunir – et de traiter de manière homogène – des phénomènes référentiels variés, dans des états de langue variés, afin de permettre des comparaisons *a posteriori*. En effet, le groupe MC4[1] est composé de linguistes de formations et de domaines de spécialité variés : certains s'intéressent plutôt à la syntaxe, d'autres à la sémantique, certains ont pour objet d'étude l'ancien français, avec une approche diachronique, d'autres le français moderne, etc. Le corpus constitué est le reflet de cette diversité.

Face à l'hétérogénéité du corpus se pose la question de la structure d'annotation : est-il possible de trouver une grille d'annotation commune à l'ensemble des textes ? Comment procéder pour annoter un tel corpus ?

2.2. La grille d'annotation

Avant la phase d'annotation proprement dite, nous avons procédé à une phase de test avec comme but l'établissement d'une grille commune d'annotation. La figure 1 restitue le jeu d'étiquettes – types d'annotation, avec pour chacun d'entre eux des propriétés et des valeurs associées à chaque propriété – qui a été mis en place et qui constitue le schéma d'annotation, obtenu au prix de réajustements permanents pendant la phase de test.

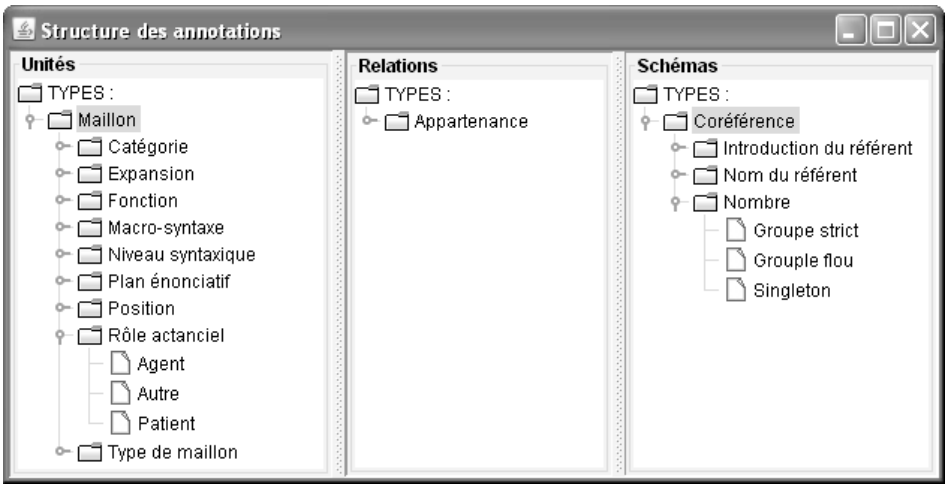

Figure 1. Grille d'annotation pour l'ensemble du corpus MC4.

Ainsi la grille a évolué au fil du projet, afin de correspondre au mieux aux préoccupations des chercheurs et aux divers types de texte. En effet, les différents états de langue ont nécessité :

- l'extension de la notion de *maillon* à ce qui n'a pas de trace linguistique marquée du fait de phénomènes d'ellipse ou de grammaticalisation. Il s'agit notamment des sujets zéro, des verbes à l'infinitif ou au participe, et particulièrement de l'élision du pronom sujet en français médiéval ;
- l'augmentation du nombre de valeurs possibles. Il est nécessaire de pouvoir repérer et annoter des cas tels que *les chevaliers le roi* (par comparaison avec *les chevaliers du roi*).

L'approche contrastive, qui met en rapport français médiéval et français contemporain, a pris tout son sens : elle a permis de mieux appréhender ces phénomènes et a été prise en compte

[1] Le projet MC4 a bénéficié de trois contrats de vacation. Nous tenons à remercier les trois vacataires qui ont participé à la procédure d'annotation, des tests initiaux à l'annotation finale.

dans les modélisations. La figure 1 donne une idée de l'étendue de cette grille d'annotation, seules les différentes valeurs possibles des propriétés « rôle actanciel » et « nombre » étant visibles.

2.3. Le travail d'annotation

L'annotation se fait en plusieurs étapes bien définies. Ces étapes ont évolué au fil des expériences d'annotation, elles sont le résultat d'une réflexion sur le travail mené et la méthodologie à adopter.

La première étape du travail est l'annotation des phénomènes référentiels. Elle se subdivise en deux temps. Premièrement, il s'agit de délimiter les maillons. C'est la seule unité d'annotation que nous ayons utilisée pour étiqueter les textes. Deuxièmement, il s'agit d'attribuer à chacun des maillons identifiés un ensemble de propriétés[2]. Chacune d'elles possède une liste de valeurs définies et décrites dans un manuel d'annotation, soit un ensemble de 78 valeurs. L'annotateur choisit pour chacune des propriétés une et une seule valeur, comme l'illustre la figure 2. Chaque maillon est ainsi défini par 11 caractéristiques, retenues parmi les 78 mises à disposition dans les listes de choix.

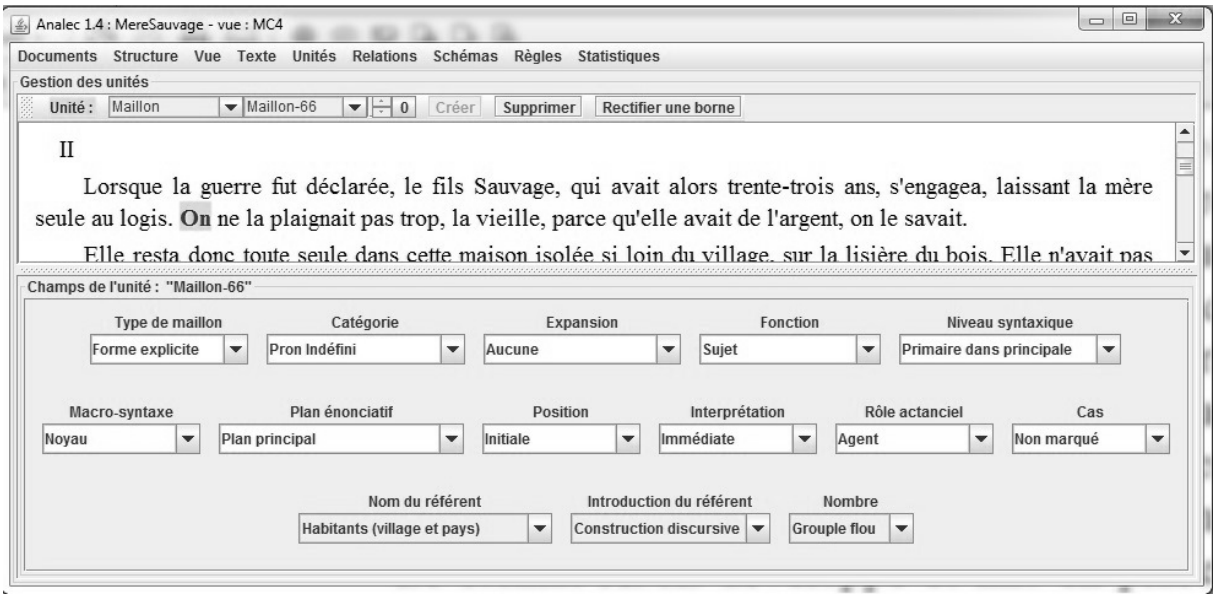

Figure 2. Annotation d'un maillon.

Nous reprenons ci-après chacune des propriétés du schéma d'annotation. A titre informatif figure entre parenthèses à côté du nom de la propriété le nombre de valeurs constitutives de la liste de valeurs de la propriété. A titre d'exemple, seule est reprise et succinctement commentée la valeur attribuée au maillon « on » présent au début du second chapitre de *La mère Sauvage* (cf. figure 2).

- *type de maillon* (2) : le référent est désigné de manière *explicite*[3] (*via* le marqueur « on », marqueur qui est délimité dans l'outil et apparaît ainsi avec un surlignement) ;
- *catégorie* (26) : « on » est un *pronom indéfini* ;
- *expansion* (11) : « on » n'a pas d'expansion[4] ;

[2] Cette méthodologie a été mise en place et empruntée par l'ensemble des annotateurs. Nous revenons sur cette pratique en section 5.1.

[3] Ce champ a deux valeurs possibles : *forme explicite* ou *forme atténuée*.

- *fonction* (13) : « on » est sujet du verbe « plaindre » ;
- *niveau syntaxique* (8) : « on » est dans une proposition principale, au premier niveau syntaxique ;
- *macro-syntaxe* (3) : « on » appartient au *noyau* prédicatif ;
- *plan énonciatif* (3) : « on » est sur le *plan principal* du point de vue énonciatif ;
- *position* (5) : « on » est en position *initiale* dans la phrase ;
- *interprétation* (2) : « on » est interprétable *immédiatement* ;
- *rôle actanciel* (3) : « on » est *agent* ;
- *cas* (2) : « on » est *non marqué*[5].

La propriété *type de maillon* donne une qualification référentielle générale à l'unité délimitée. Elle est obligatoirement renseignée par l'annotateur. L'ensemble des autres propriétés n'est pas obligatoirement informé ; s'il s'agit d'un référent de *forme atténuée*, seule la propriété *catégorie* sera informée. L'attribution d'une valeur à une propriété peut entraîner l'attribution d'autres valeurs pour d'autres propriétés : un *déterminant* (valeur de la propriété *catégorie)* aura toujours pour valeur *sans expansion* (propriété *expansion*) et *deuxième niveau* (propriété *niveau syntaxique*).

Comme le montre le tableau 1, le corpus annoté MC4 contient 4066 maillons annotés, ce qui représente 44726 choix de valeurs. Ce chiffre donne une idée de l'effort d'annotation réalisé. Il est par ailleurs intéressant de noter qu'on trouve en moyenne un maillon d'une CR tous les cinq mots d'un texte. Afin d'optimiser cette phase du travail d'annotation, nous avons été amenés à réfléchir sur des aspects ergonomiques et à envisager diverses solutions techniques, par exemple :

- minimiser le nombre de clics souris ;
- penser l'ergonomie du formulaire d'annotation (répartition horizontale et verticale des propriétés, ordonnancement, espace entre chaque propriété, etc.) ;
- préparer un éventuel remplissage automatique de certaines catégories.

La seconde étape est l'annotation des phénomènes de coréférence. Les structures à annoter que sont les CR couvrent le texte du début à la fin. Les étapes de travail sont les suivantes :

- repérer un personnage et matérialiser la chaîne correspondante ;
- annoter la chaîne (manuellement ou automatiquement) : attribuer les valeurs intrinsèques à la coréférence ;
- annoter ou récupérer des données extrinsèques pour autoriser des calculs de corrélations – pas seulement les marques de changement de paragraphe, cela peut être aussi des annotations à la ANNODIS, cf. (Pery-Woodley *et al*., 2009) ;
- visualiser et interroger les données annotées.

Trois champs sont dédiés à l'annotation des CR. Ils apparaissent sur la dernière ligne de la figure 2 (*supra*). Ils permettent d'indiquer :

- le *nom du référent* : « habitants (village et pays) » est une étiquette qui permet d'identifier et nommer de manière unique le référent ;

---

[4] Les valeurs de ce champ ne sont pas binaires (de type *oui* ou *non*), comme pourrait le laisser croire l'exemple développé. Si le maillon a une expansion, il s'agit ici de définir le type d'expansion : complément du nom, apposition, etc.

[5] La propriété *cas* est utile et nécessaire pour annoter les textes en ancien français.

- l'*introduction du référent* : « construction discursive » signifie qu'il n'est pas fait mention du référent « habitants » en amont dans le texte. L'étiquette caractérise le contexte dans lequel apparaît le maillon ;
- le *nombre* : « groupe flou » signifie qu'il s'agit d'un ensemble de personnes, sans qu'il soit possible de préciser qui sont exactement ces personnes, qui appartient ou n'appartient pas au groupe (Landragin, 2011 ; Landragin & Tanguy, ce volume).

A titre indicatif, le corpus annoté MC4 contient 285 CR. Certaines chaînes ne comportent que trois ou quatre maillons. D'autres en comportent beaucoup plus, par exemple 109 pour le personnage de Sylvie Fabre dans la nouvelle d'Echenoz – total obtenu en tenant compte de l'ensemble des paragraphes du texte, donc en incluant toutes les sous-chaînes au sens de (Schnedecker, 1997). Ces écarts importants dans les longueurs de chaînes, ainsi, bien entendu, que la taille modeste des extraits annotés, expliquent le rapport entre nombre de maillons et nombre de chaînes : dans notre corpus, une CR comporte en moyenne 14 maillons.

Face au corpus constitué, aux annotations à mettre en place, à la quantité de données à analyser, quelles sont les possibilités et moyens dont dispose le linguiste ? Comment peut-il utiliser et enrichir le corpus ?

3. Un exemple de linguistique de corpus outillée

3.1. Les besoins du linguiste

En s'appuyant sur le corpus, le linguiste mène un travail de réflexion sur la langue. Il observe en contexte des phénomènes linguistiques. Face à un corpus (textes et annotateurs multiples) et ce que l'on souhaite en tirer, se pose très vite, dès le début du projet, la question de l'outil : comment appréhender le corpus ? Quel outil de travail choisir ? D'un point de vue « pratique » : peut-on envisager de faire de la linguistique de corpus avec un logiciel de traitement de texte tel que Word ?

Le linguiste peut importer dans un logiciel de traitement de texte, le ou les textes qu'il souhaite annoter. Ce type d'outil permet l'informatisation, aussi minimale soit-elle, du corpus. C'est un moyen parmi d'autres pour appliquer sur un texte des annotations, charge à l'annotateur d'associer à une annotation une mise en forme dans le traitement de texte (par exemple : grasseyer les maillons, associer à chacune des valeurs de chacune des propriétés une couleur, surligner chacun des maillons avec la couleur adéquate). Mais ce n'est évidemment pas le moyen le plus approprié et le plus pérenne. Si l'import des données y est aisé et immédiat, ce type d'outil atteint rapidement ses limites. En utilisant un tel logiciel, le linguiste peut difficilement traiter un grand nombre de données (annotation quantitative), peut difficilement annoter avec précision et finesse (annotation qualitative). L'outil ne l'aide pas à mener une analyse tant quantitative que qualitative des données.

De la constitution du corpus à l'annotation, en passant par la mise en place de la grille d'annotation, il est nécessaire au fil du travail effectué de pouvoir analyser et rectifier les annotations, rectifications qui peuvent être plus ou moins profondes et nécessaires. Si l'aspect chronologique des étapes énumérées ci-dessus est tangible, chacune d'entre elles doit pouvoir être réitérée, être déplacée sur cet axe chronologique et monopolisée à tout moment du travail. Le travail d'annotation est en effet cyclique, il reboucle sur certaines des étapes mentionnées. L'outil est une aide : il facilite et optimise le travail d'annotation, en tenant compte du caractère cyclique de ce dernier.

Le travail d'annotation de corpus demande en amont d'inventorier un certain nombre d'outils existants et de s'interroger sur :

- l'adéquation outil-annotateur. Il s'agit entre autres d'apprécier l'ergonomie et la convivialité d'un outil, ainsi que sa prise en main, qui doit être facile pour chacun des membres du projet. Il s'agit aussi de s'assurer qu'au moins un des membres du projet en a l'expertise (ou peut tout au moins l'acquérir) ;
- l'adéquation outil-tâche. Quels sont les atouts, les limites et les manques d'un outil face aux phénomènes à étudier, aux hypothèses de travail à tester ? Peut-on pallier ces manques ? Est-il possible de faire évoluer un outil existant, ou faut-il envisager de concevoir un nouvel outil entièrement dédié à l'annotation et l'interprétation de phénomènes référentiels ?

Les annotations sont une valeur ajoutée au(x) texte(s) qu'il convient de valoriser et pérenniser. Pour cela, indépendamment d'une interface d'annotation spécifique, il est nécessaire de s'intéresser aux formats de sauvegarde du corpus annoté. C'est en effet le ou les fichiers obtenus en sortie de l'outil d'annotation qui vont ensuite être utilisés pour la diffusion et l'exploitation du corpus dans la communauté scientifique.

3.2. La valorisation du corpus

XML est un format standard d'échange et de structuration des données. Il permet d'enrichir un texte avec des annotations, et ce de manière rationnelle, en suivant des principes de balisage et de codage des données. Nous décrivons ci-après un format XML possible pour l'annotation des maillons et des CR. Ce format intègre une structuration XML de notre schéma d'annotation et suit les principes de la TEI (*Text Encoding Initiative*, cf. http://www.tei-c.org/). Il est produit automatiquement par l'outil[6] à partir des annotations réalisées, ce qui permet l'exploitation automatisée de celles-ci en même temps que leur sauvegarde dans un format de fichier ouvert. A titre d'illustration, nous prenons les deux premières phrases du second chapitre de *La mère Sauvage* de Maupassant (cf. section 2.1).

Le format XML – reproduit dans la figure 3 – est constitué de balises textuelles, <*p*> et </*p*>, qui englobent les unités de type paragraphe, et de balises délimitant les annotations. Dans le paragraphe considéré, dix maillons sont repérés, numérotés de 61 à 70. Ainsi « la mère », en gras dans l'exemple, est le maillon 65. Chaque maillon est délimité par des ancres <*anchor*>, balise qui possède les attributs suivants : *xml:id*, *type* et *subtype*. Le premier assigne un numéro unique (qui servira à retrouver plus loin les annotations), le second qualifie le type d'élément annoté, le troisième détermine s'il s'agit du début ou de la fin du maillon. L'utilisation d'ancres permet de gérer des unités qui se superposent, s'enchâssent ou se chevauchent. Ainsi le pronom relatif « qui », maillon 62, est enchâssé dans le maillon 61, ce que l'on voit dans le texte de la figure 3.

[6] Bien que les extraits présentés soient issus d'ANALEC, on peut souligner que tout outil d'annotation devrait être capable d'exporter dans un format XML de ce type.

```
<p>
Lorsque la guerre fut déclarée, <anchor xml:id="u-Maillon-61-start"
type="AnalecDelimiter" subtype="UnitStart"/>le fils Sauvage, <anchor xml:id="u-
Maillon-62-start" type="AnalecDelimiter" subtype="UnitStart"/>qui<anchor xml:id="u-
Maillon-62-end" type="AnalecDelimiter" subtype="UnitEnd"/> avait alors trente-trois
ans<anchor xml:id="u-Maillon-61-end" type="AnalecDelimiter" subtype="UnitEnd"/>,
<anchor xml:id="u-Maillon-63-start" type="AnalecDelimiter"
subtype="UnitStart"/>s&apos;<anchor xml:id="u-Maillon-63-end"
type="AnalecDelimiter" subtype="UnitEnd"/>engagea, <anchor xml:id="u-Maillon-64-
start" type="AnalecDelimiter" subtype="UnitStart"/>laissant<anchor xml:id="u-
Maillon-64-end" type="AnalecDelimiter" subtype="UnitEnd"/> <anchor xml:id="u-
Maillon-65-start" type="AnalecDelimiter" subtype="UnitStart"/>la mère<anchor
xml:id="u-Maillon-65-end" type="AnalecDelimiter" subtype="UnitEnd"/> seule au
logis. <anchor xml:id="u-Maillon-66-start" type="AnalecDelimiter"
subtype="UnitStart"/>On<anchor xml:id="u-Maillon-66-end" type="AnalecDelimiter"
subtype="UnitEnd"/> ne <anchor xml:id="u-Maillon-67-start" type="AnalecDelimiter"
subtype="UnitStart"/>la<anchor xml:id="u-Maillon-67-end" type="AnalecDelimiter"
subtype="UnitEnd"/> plaignait pas trop, <anchor xml:id="u-Maillon-68-start"
type="AnalecDelimiter" subtype="UnitStart"/>la vieille<anchor xml:id="u-Maillon-68-
end" type="AnalecDelimiter" subtype="UnitEnd"/>, parce qu&apos;<anchor xml:id="u-
Maillon-69-start" type="AnalecDelimiter" subtype="UnitStart"/>elle<anchor
xml:id="u-Maillon-69-end" type="AnalecDelimiter" subtype="UnitEnd"/> avait de
l&apos;argent, <anchor xml:id="u-Maillon-70-start" type="AnalecDelimiter"
subtype="UnitStart"/>on<anchor xml:id="u-Maillon-70-end" type="AnalecDelimiter"
subtype="UnitEnd"/> le savait.
</p>
```

Figure 3. Balisage XML du texte annoté.

Au-delà du texte et des repérages à l'aide d'ancres, le fichier XML dresse la liste des annotations. Un maillon correspond à une unité annotée. Prenons le maillon 65, « la mère ». La figure 4 montre comment ce maillon, construit à partir de l'ancre initiale et de l'ancre finale, est défini avec l'identifiant : *u-Maillon-65* (*u* indiquant qu'il s'agit d'une unité). Une fois le maillon défini, il s'agit de lui associer les annotations correspondantes. La figure 5 montre les 11 propriétés définies plus haut (2.2), spécifiées ici sous la forme d'une structure de traits.

```
<span xml:id="u-Maillon-65"
from="#u-Maillon-65-start" to="#u-Maillon-65-end"
ana="#u-Maillon-65-fs"/>
```

Figure 4. Balisage XML d'un maillon.

```
<fs xml:id="u-Maillon-65-fs">
<f name="Cas"><string>Non marqué</string></f>
<f name="Fonction"><string>Complément Direct</string></f>
<f name="Plan_énonciatif"><string>Plan principal</string></f>
<f name="Macro-syntaxe"><string>Suffixe</string></f>
<f name="Interprétation"><string>Immédiate</string></f>
<f name="Rôle_actanciel"><string>Patient</string></f>
<f name="Catégorie"><string>GN Défini</string></f>
<f name="Position"><string>Médiane</string></f>
<f name="Expansion"><string>Aucune</string></f>
<f name="Niveau_syntaxique"><string>Primaire dans enchâssée</string></f>
<f name="Type_de_maillon"><string>Forme explicite</string></f>
</fs>
```

Figure 5. Balisage XML des propriétés et valeurs d'un maillon.

Le maillon 65, au même titre que de nombreux autres dans le texte, désigne un personnage, « la mère ». Comme le montre la figure 6, la balise *<join>* permet de lister les maillons ayant le même référent. Les maillons désignant la mère étant nombreux, nous avons tronqué le contenu de la balise (*[...]*). La liste est définie comme chaîne de référence, avec dans notre

exemple l'identifiant : *s-Coréférence-14*. Le *s* indique qu'il s'agit d'un schéma (Widlöcher *et al.*, 2009), d'un ensemble de maillons. Chaque maillon est listé comme unité constitutive de la CR de la manière suivante : *#u-maillon- + identifiant du maillon*. Enfin, cette CR est annotée avec ses propres propriétés, ce que l'on peut voir – ici aussi sous la forme d'une structure de traits – dans la figure 7.

```
<join xml:id="s-Coréférence-14"
target="#u-Maillon-303 #u-Maillon-424 #u-Maillon-336 #u-Maillon-65 #u-Maillon-267
#u-Maillon-78 #u-Maillon-67 […] ana="#s-Coréférence-14-fs"/>
```

Figure 6. Balisage XML d'une chaîne de référence.

```
<fs xml:id="s-Coréférence-14-fs">
<f name="Introduction_du_référent"><string>Association</string></f>
<f name="Nombre"><string>Singleton</string></f>
<f name="Nom_du_référent"><string>La mère Sauvage</string></f>
</fs>
```

Figure 7. Balisage XML des propriétés et valeurs d'une chaîne de référence.

Ainsi, les figures 3 à 7 donnent une idée d'un possible format de représentation d'un corpus annoté en CR. Face aux besoins du linguiste et aux données à formaliser et analyser, quel outil et quelles solutions techniques sont nécessaires et disponibles ? De fait, et c'est ce que nous allons voir maintenant, les extraits XML présentés ont été générés par l'outil ANALEC qui a servi tout au long du projet MC4.

4. L'outil ANALEC

4.1. Un outil adapté à l'annotation de phénomènes référentiels

Dans le cadre du projet MC4, nous avons choisi d'utiliser ANALEC[7] (Landragin *et al.*, 2012), préexistant au projet mais alors en cours de développement, suite au constat qu'il n'existait pas d'outil adapté à l'annotation manuelle des expressions référentielles et des CR, avec prise en compte d'une évolution possible de la grille d'annotation en cours de projet. Ainsi, l'objectif du projet MC4 est double. Non seulement, il s'agit de mettre en place une méthodologie d'annotation pour les éléments référentiels : identifier et résoudre les problèmes d'annotation, obtenir un corpus annoté finement. Mais il s'agit aussi de faire évoluer l'outil utilisé afin d'optimiser les analyses linguistiques et de faciliter le travail du linguiste.

Nous illustrons ci-après comment ANALEC a su évoluer et prendre en considération les besoins exprimés par les annotateurs et les chercheurs, au travers de son interface (section 4.2), des formats d'échange standards (section 4.3), de l'adaptabilité de la grille d'annotation (section 4.3) et de fonctionnalités d'analyse quantitative et visuelle des données (section 4.4), cette dernière étape concernant à la fois les annotateurs et les chercheurs, quand il ne s'agit pas des mêmes personnes, bien entendu.

4.2. Une interface optimisée

Le texte est présenté dans la moitié supérieure de la fenêtre principale, et l'interface graphique d'annotation – le formulaire – dans sa moitié inférieure. L'utilisateur peut ainsi délimiter les unités dans le texte et leur affecter un type ainsi que les valeurs associées en peu de

[7] Le logiciel dans sa version 1.4, c'est-à-dire celle issue du projet MC4, peut être téléchargé à l'adresse suivante : http://www.lattice.cnrs.fr/Telecharger-Analec

manipulations (Landragin *et al*., 2012). L'essentiel du travail consiste de fait à choisir des valeurs dans des menus déroulants.

Comme de nombreux outils, ANALEC met à disposition de l'annotateur des raccourcis clavier. Ils ont été instaurés à la demande des annotateurs, répondant ainsi à des besoins exprimés. Certains sont conventionnels et utilisés dans de nombreuses interfaces classiques. C'est le cas de *Ctrl+S* qui permet d'enregistrer à tout moment le travail effectué. Quand une étoile apparaît à la suite du nom du fichier en haut à gauche de la fenêtre, cela signifie que les dernières modifications apportées ne sont pas enregistrées. Par ailleurs, dans le formulaire d'annotation, la tabulation permet de passer d'un champ à un autre. Les champs sont placés et hiérarchisés dans le formulaire grâce à la « vue », notion permettant la mise en forme et l'accès aux différentes données (Landragin *et al*., 2012). De manière plus spécifique, les touches F1 et F2 permettent de passer d'une unité annotée à l'autre, au fil du texte, en allant vers le bas pour la première et vers le haut pour la seconde. L'utilisation de F1 et F2 permet de naviguer très efficacement, non seulement dans les CR, mais également dans les annotations des maillons : en se focalisant sur une propriété, on peut faire défiler les valeurs prises par les différents maillons annotés dans le texte.

4.3. Des formats d'échange standards

ANALEC permet l'import et l'export de données en divers formats. Il est possible d'importer du texte brut (*.txt*) ou des textes déjà annotés, ces derniers devant être soit au format GLOZZ (Widlöcher & Mathet, 2009) – qui dissocie texte (*.ac*) et annotation (*.aa*) –, soit au format XML dans un fichier unique, comme détaillé dans la section 3.2.

Une fois le texte importé dans l'outil, il est possible de nettoyer et de modifier le texte dans l'interface du logiciel. Cette action peut être menée à tout moment du travail : quand le texte est encore brut ou en cours d'annotation – ce qui est assez rare parmi les outils d'annotation existants pour être mentionné. ANALEC remplit ainsi pleinement sa fonction d'éditeur de texte, l'intérêt étant de pouvoir corriger une erreur dans le texte sans perdre les annotations déjà saisies. On peut bien sûr s'interdire ce genre d'action, notamment quand le corpus est partagé entre plusieurs annotateurs dans un but de comparaison.

Dans une certaine mesure, il est possible de passer de Word à ANALEC et d'ANALEC à Word grâce à des scripts (petits programmes informatiques faciles à écrire) développés dans le cadre de ce projet et dans d'autres projets réalisés avec ANALEC. Un script prend par exemple en entrée des fichiers *.doc* et génère en sortie des fichiers *.aa* et *.ac* en interprétant les surlignements effectués dans Word. Ceci permet à certains linguistes de continuer à travailler et annoter avec leur traitement de texte habituel, du moins avec certaines contraintes (codage couleur d'une seule propriété, par exemple). Rappelons ici que le groupe de travail MC4 est essentiellement constitué de linguistes, et non d'informaticiens. Le principal du travail d'annotation effectué est manuel. Les outils, logiciel et scripts, viennent en aide à l'annotateur et allègent tant que faire se peut les tâches humaines. Ces derniers sont aussi une façon de montrer qu'un premier travail effectué dans Word peut être rationalisé ensuite dans un outil dédié à l'annotation.

Concernant les échanges avec d'autres outils voire des outils à venir, tout repose sur le format XML encodant texte et annotations. Comme nous l'avons vu dans la section 3.2, ce format XML, développé et mis en place spécifiquement pour les annotations des CR du projet MC4, répond aux normes du consortium TEI, ce qui garantit sa pérennité et son utilisabilité.

4.4. L'adaptabilité de la grille d'annotation

Une étape du projet MC4 a consisté à mettre en place une grille d'annotation. Sous ANALEC, cette grille se nomme « structure des annotations » (cf. figure 1) et a deux atouts non négligeables. Elle peut être partagée par un groupe, tout en étant évolutive :

- Partager une structure d'annotation : la structure d'annotation contient comme nous l'avons mentionné un grand nombre d'étiquettes. Au seul type « maillon » sont associées 11 propriétés et 78 valeurs[8]. Notons cependant que 5 types ont une valeur par défaut (indiquée entre parenthèses après chacun des types) qui est renseignée dès la création d'un maillon : cas (*non marqué*), expansion (*aucune*), macro-syntaxe (*noyau*), niveau syntaxique (*principale*) et type de maillon (*forme explicite*). Cette valeur est celle qui est la plus souvent attribuée. Elle est renseignée automatiquement mais elle peut bien entendu être modifiée par l'annotateur. Grâce à la notion de vue propre à ANALEC, un groupe peut accéder à un formulaire commun d'annotation, formulaire convivial où les nombreux champs et valeurs d'annotation peuvent être ordonnés selon les besoins des utilisateurs. Des effets visuels comme la colorisation d'éléments annotés participent à la convivialité de l'interface graphique.
- Faire évoluer la structure d'annotation : l'atout principal d'ANALEC est très certainement de permettre au travail de recherche d'évoluer. Le schéma d'annotation n'a pas à être défini – ou plutôt à être définitif – en début de projet. Il peut évoluer pendant la phase d'annotation. Ce caractère dynamique d'ANALEC est conforme aux aléas fréquents d'un travail de recherche, et c'est lui qui a orienté notre choix vers cet outil pour ce projet.

Il peut être utile de veiller à différencier les droits de chacun des annotateurs : certains d'entre eux réfléchissent sur le schéma d'annotation, le font évoluer, le déclinent en plusieurs versions (une version allégée peut côtoyer une version très complète, avec des traits temporaires par exemple), d'autres ne doivent rien modifier. La vue, avec les contraintes qu'elle permet d'imposer ou de relâcher, est de ce fait une solution technique efficace.

A partir du moment où plusieurs versions d'un schéma d'annotation existent et sont utilisées, l'outil doit permettre de gérer ces versions de manière rationnelle. Avec ANALEC, il est ainsi possible de fusionner plusieurs schémas d'annotation en un seul. En partant d'un corpus annoté, ceci permet de faire correspondre les annotations déjà réalisées avec un autre schéma d'annotation. En fonction des cas de figure, les propriétés et valeurs seront regroupées dans le nouveau schéma. L'utilisateur garde le contrôle des fusions et des regroupements, de même qu'il peut aisément renommer propriétés et valeurs pour rendre un schéma compatible avec un autre. Renommer permet aussi de rattraper des erreurs en regroupant plusieurs variantes orthographiques de ce qui constitue en fait une même valeur ou une même propriété.

4.5. Une analyse quantitative et visuelle des données

ANALEC permet d'obtenir des analyses quantitatives et visuelles, soit des unités (les maillons pour le projet MC4), soit des schémas (les CR).

Une analyse quantitative des unités est accessible *via* l'onglet « statistiques ». L'annotateur accède, pour l'ensemble des valeurs des propriétés de chaque unité, à des calculs de fréquences, des calculs de corrélations et à des analyses factorielles de correspondance *via* des représentations géométriques sous la forme de nuages de points. L'outil permet ainsi à tout moment d'observer les données dans leur ensemble. Il offre une vision du travail

[8] Le manuel d'annotation qui, par manque de place, ne peut être reproduit ici, sera mis à disposition avec le corpus étiqueté lors de la diffusion de celui-ci.

effectué, le calcul du nombre d'étiquettes attribuées et des catégories associées. Le linguiste peut se faire rapidement une idée sur son intuition initiale, afin de faire évoluer ou non la structure (ou encore le corpus, par exemple). Il est possible non seulement de faire des décomptes sur un texte, mais aussi sur un ensemble de textes. C'est le rôle de la fonction « concaténation », qui consiste à regrouper des fichiers annotés à l'aide du même schéma d'annotation.

Dans les formulaires ANALEC, il n'y a pas de champs conditionnels. A savoir, une valeur incrémentée dans un champ n'influe pas le remplissage d'un autre champ, par exemple selon des règles du type : si un champ est rempli, il faut obligatoirement remplir tel autre champ, ou : si un champ a telle valeur, alors tel autre champ a une liste de valeurs restreintes. Ce type de règles ne peut pas être défini dans les formulaires ANALEC, mais elles peuvent facilement être vérifiées dans les données annotées à l'aide du volet statistique de l'outil.

Notons que dans le cadre du projet MC4, certaines des propriétés associées à un maillon – comme par exemple la position d'une expression référentielle dans la phrase – sont aussi (voire davantage) des propriétés de phrase. Mais à des fins d'interrogation, nous les avons associées aux maillons : les calculs statistiques se font aisément dans l'ensemble des propriétés d'unités de même type, alors qu'ils sont beaucoup plus complexes à mettre en œuvre dans l'ensemble des propriétés d'unités de types différents, comme ce serait le cas avec des unités pour les maillons et des unités pour les phrases. Tout regrouper dans les unités « maillon » permet – en contournant d'une certaine manière les limites de l'outil – d'interroger efficacement les données *via* le volet statistique.

Le module de visualisation des chaînes, quant à lui, permet de visualiser la répartition des unités et des propriétés dans le texte, en suivant l'ordre linéaire des annotations. Il a spécifiquement été développé pour le projet MC4, pour la visualisation des CR. Il permet – *via* un code couleur et une représentation graphique simple à base de points – de visualiser chacune des propriétés annotées (cf. figure 8).

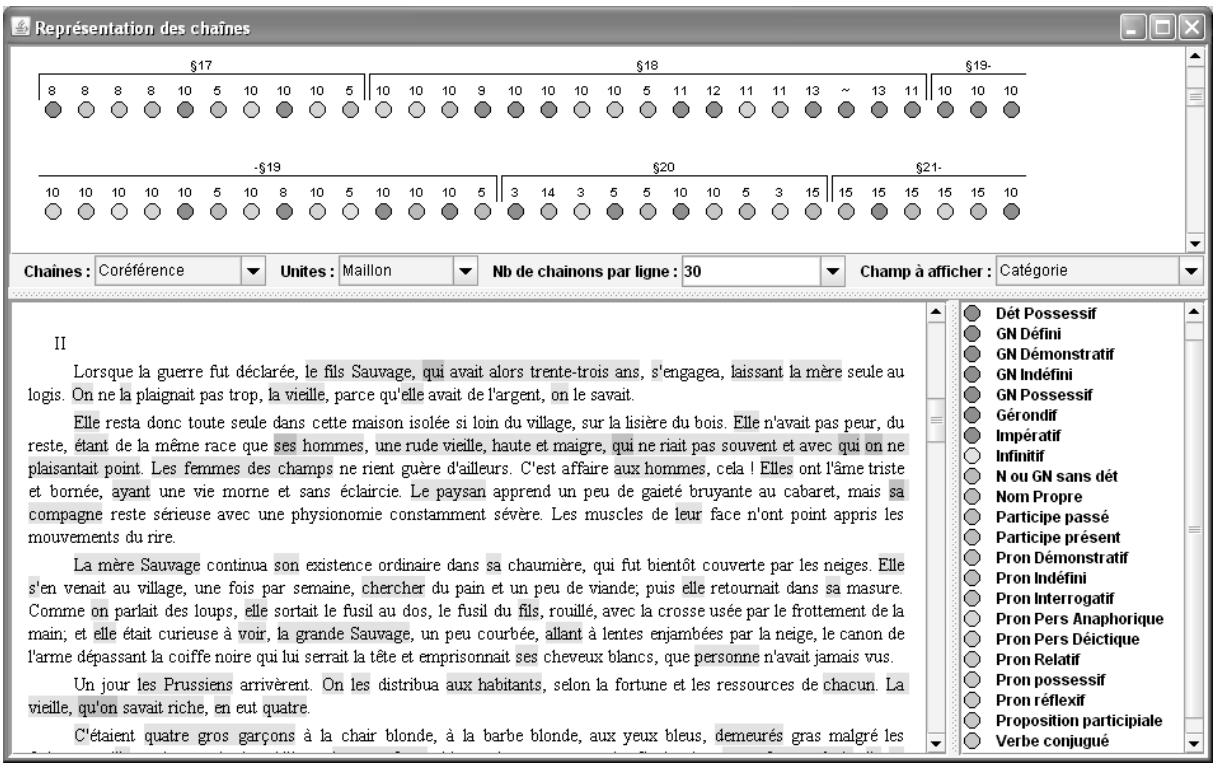

Figure 8. Copie d'écran de l'interface d'ANALEC : visualisation de l'ensemble des chaînes de référence de *La mère Sauvage* (Maupassant).

Dans la copie d'écran de la figure 8, nous visualisons, à partir du dix-septième paragraphe du texte, l'ensemble des maillons annotés et la catégorie associée à chacun de ces maillons. A une catégorie est associé un code couleur (reproduit ici en nuances de gris), repris en haut de la fenêtre dans la représentation graphique des CR. L'important est de voir que les codes couleur, s'ils permettent d'appréhender le matériau linguistique d'une manière graphique assez agréable, ont néanmoins leurs limites. Rappelons que l'œuvre de Maupassant contient plus de 500 maillons et que les catégories sont au nombre de 26, il est donc difficile d'observer un quelconque phénomène quand on visualise sur la CR l'ensemble des informations disponibles.

En revanche, si nous sélectionnons un référent spécifique – par exemple « la mère sauvage », et que nous projetons sur la chaîne un sous-ensemble des catégories (par exemple *GN défini* et *Pronom Personnel Anaphorique*), alors la CR épurée est plus lisible. C'est ce que montre la figure 9, où l'on peut commencer à analyser l'alternance des maillons clairs (groupes nominaux définis) et des maillons foncés (pronoms anaphoriques de troisième personne), à analyser la fréquence d'apparition des maillons clairs en début de paragraphe, etc. ANALEC permet de sélectionner des sous-ensembles de valeurs de manière cumulative, pour une même propriété, mais aussi en parcourant les propriétés l'une après l'autre. Il est ainsi facile d'obtenir une représentation graphique des catégories de maillons, des fonctions syntaxiques, des rôles actanciels, pour l'ensemble des CR comme pour un sous-ensemble, voire une seule chaîne, comme c'est le cas dans la figure 9 : on observe que le personnage de la mère apparaît pour la première fois dans le second chapitre, et ce sous la forme d'un groupe nominal défini, forme qui est souvent utilisée en début de paragraphe, du moins dans les dix premiers paragraphes où elle apparaît (le pronom est ensuite largement prépondérant).

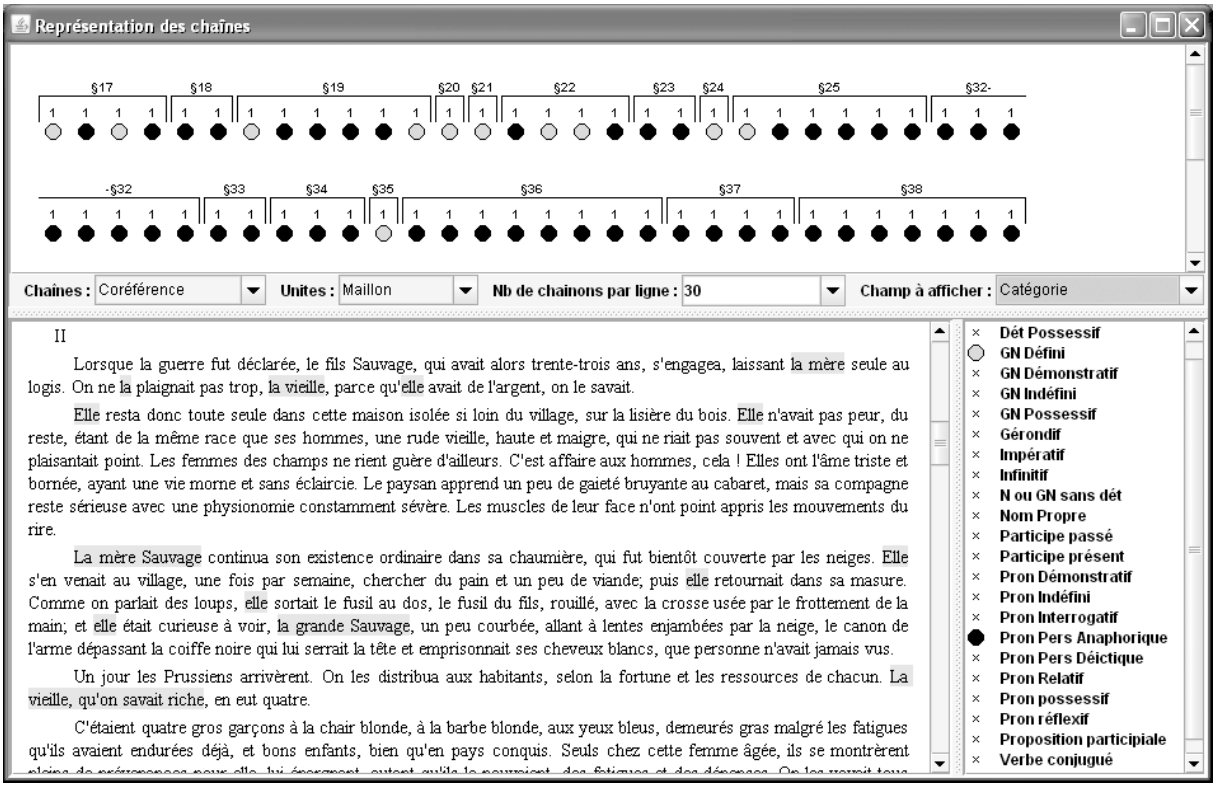

Figure 9. Copie d'écran de l'interface d'ANALEC : visualisation d'une chaîne de référence spécifique *de La mère Sauvage* (Maupassant), avec sélection de deux catégories de formes.

L'outil ANALEC regroupe donc dans une interface simple et unifiée différentes fonctionnalités, depuis l'annotation jusqu'à la visualisation des CR en passant par divers modes d'accès aux données et à des statistiques les caractérisant. Son évolution s'est faite en fonction des besoins exprimés par les utilisateurs linguistes, et notamment par des chercheurs lors de phases d'élaboration d'une grille d'annotation, avant la finalisation de celle-ci et la rédaction d'un manuel d'annotation. L'accent a été mis sur la possibilité, à tout moment, d'une part de revenir sur des choix sans perdre le travail déjà effectué (annotation dynamique), d'autre part de visualiser les données annotées, notamment pour détecter des erreurs ou des exemples remarquables (boucle visualisation-annotation). L'utilisation d'ANALEC dans le cadre du projet MC4 a permis d'aller plus loin dans l'étude des CR, et aussi de faire apparaître de nouveaux besoins. C'est ce que nous allons détailler dans la section suivante.

5. ANALEC, un outil à adapter en permanence

5.1. Une annotation manuelle dynamique

Notre groupe de travail MC4 a pu annoter, visualiser, faire des calculs. Grâce à la linguistique de corpus outillée et plus particulièrement à ANALEC, les linguistes sont allés plus loin, ont pu vérifier certaines de leurs intuitions dans l'étude de phénomènes de coréférence. De manière réciproque, grâce au travail des linguistes, ANALEC a évolué et a pris en considération les besoins spécifiques des annotateurs. Il est ainsi devenu un outil utile et pertinent pour l'annotation, l'analyse et la visualisation des CR[9].

Le travail d'annotation du groupe MC4 a permis de faire évoluer l'outil selon trois types d'améliorations :

- corrections de bugs : les utilisateurs ont mis au jour un certain nombre de bugs qu'il s'est agi de vérifier, comprendre et corriger ;
- amélioration de l'ergonomie : les utilisateurs ont exprimé des besoins spécifiques liés à leur tâche, qui ont nécessité par exemple la mise en place de nouveaux raccourcis clavier (pour faciliter l'enregistrement et le parcours des données, par exemple), la structuration et l'agencement des champs d'annotation (enrichissement des fonctionnalités des vues, par exemple) ;
- constitution d'une procédure d'annotation : des pratiques d'annotation a découlé une chronologie des tâches d'annotation, à savoir : tout d'abord délimiter l'ensemble des maillons d'un texte ; ensuite attribuer à chacun d'eux l'ensemble des valeurs[10] ; puis repérer les CR ; et enfin attribuer à chaque CR un ensemble de valeurs.

ANALEC a été conçu pour aider des linguistes à construire leur grille d'annotation tout en annotant : un changement dans la grille ne conduit pas à la perte des annotations déjà réalisées, mais, au contraire, à leur prise en compte dans la nouvelle structure. Comme nous l'avons dit, c'est cette fonctionnalité, dite « annotation dynamique » qui a conduit les participants du projet MC4 à utiliser ANALEC plutôt qu'un outil d'annotation existant.

[9] Il ne s'agit pas de réduire ANALEC à un outil d'annotation de phénomènes de coréférence. Il est envisageable d'utiliser cet outil pour de nombreuses tâches d'annotation. Nous abordons d'ailleurs brièvement dans la section 5.4 un autre travail d'annotation mené sous ANALEC : annotation d'éléments initiaux en tête de phrase. Mais ce n'est, là encore, qu'un exemple d'utilisation de l'outil parmi d'autres.

[10] En ce qui concerne l'attribution des valeurs, les pratiques divergent : certains annotateurs préfèrent renseigner l'ensemble des champs associés à un maillon avant de passer au maillon suivant, d'autres vont renseigner un champ (ou un ensemble de champs, par exemple l'ensemble des champs relevant d'aspects syntaxiques) pour l'ensemble des maillons avant de passer au champ – ou à un ensemble de champs – suivant.

L'amélioration de l'outil pour une meilleure prise en compte des spécificités des CR était un aspect du projet MC4, et les copies d'écran présentées en figure 8 et en figure 9 sont issues de la version 1.4 d'ANALEC, c'est-à-dire la version dont le développement a suivi les commentaires et les retours d'expérience des membres du projet MC4. L'enjeu principal était la mise en place d'un module permettant la visualisation et l'annotation à partir de cette visualisation des CR. Les enjeux secondaires étaient l'amélioration – *via* une utilisation intensive – des fonctionnalités d'annotation et d'interrogation statistique de l'outil.

Notons cependant que le projet MC4 a impliqué un besoin constant de scripts – en entrée comme en sortie – pour transformer les données et optimiser le travail en amont et en aval d'ANALEC. Un exemple typique est l'écriture d'un script permettant d'extraire de l'ensemble des données annotées une sous-partie afin de réaliser un tableau chiffré illustrant un article de recherche. Tout cela est coûteux en temps, et montre le manque d'immédiateté de l'outil (et, de fait, de n'importe quel outil). ANALEC peut donc encore évoluer sur cet aspect de valorisation, de même que sur ceux plus techniques que nous abordons maintenant.

5.2. L'annotation de corpus et la gestion de base de données

Une refonte d'ANALEC afin de travailler non plus avec un corpus annoté mais avec une véritable base de données serait nécessaire, pour mieux gérer la notion de textes constitutifs d'un corpus, par exemple, et pour exploiter de manière plus efficace les procédés classiques d'interrogation de bases de données. En effet, s'il est possible de concaténer des textes dans ANALEC pour se constituer un « gros » corpus, on perd alors du même coup la notion de texte, et la possibilité de comparer les annotations texte par texte. Le manque peut actuellement être pallié par l'utilisation de scripts, par exemple en PERL : langage fréquemment utilisé pour formater des données (Tanguy & Hathout, 2007). Pour ce faire, il est nécessaire d'exporter les données du format ANALEC (*.ec*) au format GLOZZ (*.aa* et *.ac*) ou au format XML décrit.

L'intégration de bases de données dans ANALEC est une piste envisageable pour alléger et faciliter le travail. Cela permettrait dans un premier temps de structurer davantage les données (grâce à la constitution d'un schéma relationnel), pour dans un second temps en faciliter l'accès sur le *web* : mise en place d'une plateforme de consultation et d'interrogation des données, par le biais de requêtes de type XSLT par exemple.

5.3. L'exploration et l'exploitation des chaînes de référence

Les techniques graphiques d'exploration et de calcul sur les CR restent à améliorer. Le module présenté, avec les points colorés représentant chacun des maillons, est un premier pas qui permet d'explorer les propriétés annotées pour une ou plusieurs des CR d'un texte, pour un ou plusieurs types de maillons. Pour étudier plus largement les CR, on aimerait faire émerger des données sur les types de chaînes, avec des indications telles que le nombre de maillons par chaîne, la distance (en nombre de mots) d'un maillon à l'autre, etc. Le module de « représentation des chaînes » d'ANALEC ne permet pas – dans sa version 1.4 – ce type de calcul. Le manque peut actuellement être pallié par l'utilisation de scripts, mais une solution intégrée à l'outil serait bénéfique.

5.4. L'automatisation et la semi-automatisation

L'automatisation de la création de CR est une voie à explorer, en lien avec les travaux qui portent sur ce sujet en traitement automatique des langues. Si l'on fait le parallèle avec un autre projet impliquant une partie des chercheurs ayant participé à MC4 (à savoir le groupe de

travail EIOMSIT[11]), un des objectifs était – après une phase d'annotation manuelle des éléments initiaux, en début de phrase – de repérer les chaînes d'éléments initiaux. Dans ce cas de figure, une chaîne est constituée de tout ce qui est annoté et qui précède le sujet ou le verbe (si le sujet n'est pas exprimé ou s'il est postposé). Comme il faut de nombreuses manipulations sous ANALEC pour construire manuellement chacune des chaînes d'éléments initiaux, c'est une solution automatique hors ANALEC qui a été envisagée. La solution a résidé dans l'écriture d'un script permettant de construire les chaînes automatiquement. Le script repère un patron « EI+ (Sujet | Verbe) », ce qui signifie au moins une unité « élément initial », suivie du sujet ou du verbe. Reste à annoter manuellement la valeur que prend la chaîne d'éléments initiaux (Mélanie-Becquet et Prévost, *soumis*). Le script est une aide précieuse : il permet un gain de temps, et il n'oublie aucune chaîne. Mais la spécificité des CR ne permet pas ce type d'annotation automatique, en tout cas pas de manière simple (cf. l'article de présentation de ce volume).

Il manque donc pour ce type d'étude – c'est-à-dire l'étude de chaînes d'annotations – un outil unifié où tous les besoins seraient satisfaits : facilité de repérage automatique de phénomènes remarquables ; génération de graphiques adaptés ; souplesse dans la fusion d'annotations manuelles et d'annotations automatiques. Ce sont autant de perspectives pour les futures versions d'ANALEC ou pour des outils spécifiques communiquant avec ANALEC.

Conclusion

Le projet MC4 a permis d'élaborer une procédure d'annotation des phénomènes de coréférence, de tester et de faire évoluer un outil d'annotation, ANALEC, pour mieux appréhender les chaînes et aboutir à la constitution d'un corpus annoté. Certes, les failles et les manques sont patents : la procédure d'annotation présente quelques points faibles, l'outil est imparfait et reste toujours à améliorer, le corpus annoté nécessite encore un peu de nettoyage. Cependant, ce corpus a le mérite d'exister, et il a d'ores et déjà permis de mener à bien plusieurs études, comme l'a montré ce volume. Il sera mis à la disposition de la communauté, en libre accès, de même que l'est déjà ANALEC. Les questions que se sont posées les linguistes du groupe MC4, les divers points linguistiques abordés dans ce volume sont autant de champs possibles à la constitution d'un formulaire d'interrogation – sur une plateforme *web* par exemple. Si le but premier du projet n'était pas la constitution d'une telle plateforme, le travail effectué et l'homogénéité du traitement des données laisse envisageable une telle perspective. Certains projets auxquels nous avons collaboré ont donné lieu à la création de ce type d'objet – cf. (Mélanie-Becquet & Fuchs, 2011), par exemple – qui s'avère d'une grande utilité pour la communauté.

Si ANALEC a permis la constitution de ce corpus annoté, il a aussi montré ses limites : comment faire pour tirer des statistiques intéressantes ? A quels calculs faut-il penser ? De manière plus générale se pose la question, et ce pas uniquement concernant ANALEC, de l'évolution des outils d'annotation développés pour des tâches spécifiques, outils nommés *instruments* dans la terminologie de (Habert, 2005). Pour être utilisables et utilisés, les *instruments* doivent communiquer avec des *outils*, c'est-à-dire avec des « logiciels multi-usages » (*ibid.*). Il est en effet primordial qu'un *instrument* manie divers langages (requête XSLT, textes étiquetés de type TreeTagger, etc.) et permette l'importation et l'exportation de divers formats (tableau XLS/CSV, format XML, TXM, export vers R, etc.). Ainsi, dans le cas

[11] EIOMSIT (Eléments Initiaux, Ordre des Mots, Structuration Informationnelle et Textuelle) est un projet du laboratoire Lattice, mené de 2010 à 2013. Il met en place une structure d'annotation pour les éléments initiaux (Mélanie-Becquet et Prévost, *soumis*), et utilise le corpus *Chambers-Le Baron corpus of Research Articles in French* (http://ota.ahds.ac.uk/desc/2527).

de l'annotation des CR : faut-il construire un nouvel *instrument* dédié spécifiquement à cette tâche ou faut-il davantage penser à de nouvelles passerelles, à de nouveaux modules ?

Références

Habert B. (2005), « Portrait de linguiste(s) à l'instrument », *Texto!*, n° 104, http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html.

Habert B. (2005), *Instruments et ressources électroniques pour le français*. Gap/Paris : Ophrys.

Habert B., Nazarenko A., Salem A. (1997), *Les linguistiques de corpus*. Paris : Armand Colin.

Landragin, F., Poibeau, T. & Victorri, B. (2012), « ANALEC: a New Tool for the Dynamic Annotation of Textual Data. », *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turquie, pp. 357-362.

Landragin F. (2011), « Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits », *Corpus*, n° 10 : 61-80, http://corpus.revues.org.

Mélanie-Becquet F., Fuchs C. (2011), « Elaboration d'une base de données d'exemples de structures comparatives » , *Corpus*, n° 10 : 273-295, http://corpus.revues.org.

Mélanie-Becquet F., Prévost S. (soumis), « Eléments initiaux : combinaison et schémas préférentiels dans un corpus d'articles scientifiques », article soumis à la revue Corpus.

Péry-Woodley M.-P., Afantenos Stergos D., Ho-Dac L.-M., Asher N. (2011), « La ressource ANNODIS, un corpus enrichi d'annotations discursives », *Traitement Automatique des Langues*, vol. 52, n° 3 : 71-101.

Péry-Woodley M.-P., Asher N., Enjalbert P., Benamara F., Bras M., Fabre C., Ferrari S., Ho-Dac L.-M., Le Draoulec A., Mathet Y., Muller P., Prévot L., Rebeyrolle J., Tanguy L., Vergez-Couret M., Vieu L., Widlöcher A. (2009), « ANNODIS : une approche outillée de l'annotation de structures discursives », Actes de la conférence *TALN 2009*, Senlis.

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

Tanguy L., Hathout N. (2007), *Perl pour les linguistes. Programmes en Perl pour exploiter les données langagières*, Paris : Hermès Science Publications.

Widlöcher A., Mathet Y. (2009), « La plate-forme Glozz : environnement d'annotation et d'exploration de corpus », Actes de la conférence *TALN 2009*, Senlis.