L'annotation structurelle

François Daoust [1], Yves Marcoux [2], Jean-Marie Viprey [3]

[1] UQAM – Centre ATO – Québec – Canada

[2] UdeM–EBSI – GRDS – Québec – Canada

[3] UFC – Besançon – France

Résumé

Dans la tradition de l'analyse de textes par ordinateur, l'annotation et la catégorisation font partie des opérations permettant d'enrichir les données textuelles au fur et à mesure de leur analyse, éclairée par des outils statistiques et divers modes de lecture comparative. En général cependant, les unités ainsi enrichies sont des occurrences individuelles, des unités de contexte ou des formes lexicales, affublées de propriétés, attributs ou structures de traits. Mais la structuration de ces unités et leur mise en relation sont plus rarement abordées. C'est cette dimension que nous appelons l'annotation structurelle. Du point de vue de la représentation, nous proposons que l'annotation structurelle prenne la forme de documents externes d'annotation en XML respectant une syntaxe conforme aux recommandations de la Text Encoding Initiative (TEI) et s'inscrivant dans le modèle de dépôt de données adapté à la constitution de corpus de recherche. Des exemples inspirés de la linguistique textuelle seront utilisés pour illustrer cette proposition.

Abstract

In the tradition of computer-aided text analysis, annotation and categorization are among the operations used to enrich the textual material in the course of the analysis, with the help of statistical tools and various comparative reading functions. In general, however, such enrichments are applied to textual units which are single occurrences, context units, or lexical forms, and consist simply in associating properties, attributes, or feature sets to those units. The possibility of defining *structures* or *relations* among textual units is seldom considered, even though it allows a strictly larger set of enrichments to be expressible. This is what we call *structural annotation*. We propose representing structural annotations in the form of stand-off XML documents compliant with the *Text Encoding Initiative* (TEI) recommendations, and compatible with the research-corpora repository model defined in earlier work. Examples drawn from textual linguistics will illustrate our proposal.

Keywords: structural annotation, TEI, textual linguistics

1. Introduction

Dans la tradition de l'analyse de texte par ordinateur, l'annotation et la catégorisation font partie des opérations permettant d'enrichir les données textuelles au fur et à mesure de leur analyse, éclairée par des outils statistiques et divers modes de lecture comparative. En général cependant, les unités ainsi annotées sont des occurrences individuelles, des unités de contexte ou des formes lexicales affublées de propriétés, attributs ou structures de traits. Mais la structuration de ces unités et leur mise en relation sont plus rarement abordées. C'est cette dimension, que nous appelons l'annotation structurelle, que nous présenterons ici sous la forme de proposition de format de document externe d'annotation.

Nous procèderons d'abord à une mise en contexte visant à situer cette proposition dans le contexte des échanges au sein de la communauté de l'analyse des données textuelles assistée par ordinateur. Nous procéderons ensuite à la présentation d'une proposition de syntaxe concrète, XML-TEI, pour l'annotation structurelle. Pour illustrer cette proposition, nous ferons appel à une mise en forme d'exemples d'annotation structurelle tirés d'un ouvrage de Jean-Michel Adam (2005) sur la linguistique textuelle. Nous terminerons par des perspectives de recherche sur l'implantation et l'exploitation de ces structures dans un contexte d'analyse de texte par ordinateur.

2. Problématique

Nous désignons, par *annotation structurelle*, l'ajout à des ressources textuelles existantes d'annotations analytiques visant la mise en relation de segments textuels explicitant le fonctionnement de la langue, du discours et de la mise en texte. Ces mises en relation sont des pratiques de base de l'analyse textuelle dans sa tradition scolaire. Sur un plan plus formel, l'analyse syntaxique est la forme la plus connue de l'annotation structurelle avec ses forêts d'arbres qui annotent les divers composants de la proposition et de la phrase. Au-delà de la phrase, la linguistique textuelle, dans la foulée de Bakhtine (1984), perçoit le texte comme un *réseau de déterminations*.

> La linguistique textuelle a pour rôle, au sein de l'analyse de discours, de théoriser et de décrire les agencements d'énoncés élémentaires au sein de l'unité de haute complexité que constitue un texte. Elle a pour tâche de détailler les « relations d'interdépendance » qui font d'un texte un « réseau de déterminations » (Weinrich, 1973 : 174). La linguistique textuelle porte autant sur la description et la définition des différentes unités que sur les opérations dont, à tous les niveaux de complexité, les énoncés portent la trace. (Adam, 2005 : 33).

Malgré le fait que l'analyse textuelle fasse grand état des multiples structures qui traversent le texte, la tradition de l'analyse statistique des données textuelles lui a fait peu de place. Certes, plusieurs chercheurs ont situé leur travaux aux confins de l'analyse syntaxique, telle que pratiquée en traitement automatique de la langue, et de l'analyse de discours de tradition lexicométrique (voir, entre autres, Habert, 1998). Mais ces travaux sont généralement limités à la prise en compte des syntagmes nominaux dans l'analyse contrastive des énoncés. Les connexions du texte et du discours, en tant qu'*unités structuralement ouvertes* (Charolles, 1993 : 311, cité par Adam, 2005 : 36), sont rarement prises en compte.

Même si ces connexions peuvent partager le même formalisme d'annotation que les relations syntaxiques, leur nature est très différente. Adam le souligne : dès qu'on dépasse le seuil de la phrase, ce ne sont plus les *solidarités syntaxiques* qui prévalent mais plutôt « des marques et des instructions relationnelles de portée plus ou moins lointaine » (Adam, 2005 : 36). S'appuyant sur Charolles, Adam introduit l'idée de *marques instructionnelles* qui signalent au destinataire que « telle unité doit être comprise comme entretenant telle relation avec telle ou telle autre » (Charolles, 1993 : 311, cité par Adam 2005, : 36).

Dans la tradition de l'analyse statistique des données textuelles, on marque habituellement les parties du corpus. Il s'agit généralement de balisage de la structure formelle du corpus en termes de documents, de tours de parole, de locuteurs, de paragraphes, etc. Ainsi, par exemple, l'analyse factorielle des correspondances pourra, sur la base de l'analyse des fréquences lexicales de chacune des parties marquées, produire une synthèse des données contrastant simultanément les profils lexicaux et les parties du corpus. Mais ces divisions simples entre parties demeurent un pâle reflet des relations structurales entre segments textuels.

Dans la tradition de l'analyse de texte par ordinateur (ATO), certains logiciels – par exemple SATO (Daoust, 2009) – permettent d'annoter, en cours d'analyse, les unités lexicales, les occurrences et les segments afin de rendre compte d'une variété de paradigmes catégoriels. Il

reste qu'il s'agit d'une annotation *à plat* qui ne peut marquer la relation que par héritage sur les unités terminales. Ainsi, par exemple, pour marquer la relation dialogique entre locuteurs, on pourra avoir une propriété indiquant qui est l'énonciateur et une autre indiquant à qui il s'adresse. La conjonction des deux permettra de configurer dynamiquement les parties du texte et du lexique à soumettre aux analyseurs statistiques. L'annotation structurelle vise à aller au-delà de cette annotation simple, à structure implicite, en marquant sous forme de multiples graphes les connexions induites par les *marques instructionnelles* dont parle Adam. En conjonction avec le filtrage des annotations simples, le parcours des graphes permettra de contraster beaucoup plus aisément les segments textuels en fonction de leurs positions dans l'une ou l'autre des annotations structurelles.

Dans la tradition de l'ATO, la catégorisation, dans sa dimension lexicale (forme en tant que classe) et textuelle (occurrence de la forme en contexte), permet de soumettre à l'analyse statistique des fréquences de catégories marquant des résultats d'analyse et d'interprétation susceptibles, par exemple, de rendre compte d'éléments de la structure syntaxique ou sémantique de l'énoncé. L'annotation structurelle permet en plus de compter des *configurations*, c'est-à-dire des *motifs structurels* à l'intérieur de certains empans déterminés par des structures plus amples, par exemple, telle structure argumentaire dans tel type d'épisode narratif.

L'intérêt de l'annotation structurelle ne se limite pas, bien entendu, à la qualification des unités soumises au calcul statistique. Comme les concordances, par exemple, elle est un outil de navigation permettant des parcours hypertextuels appuyant l'interprétation sur l'explicitation des connexions qui tissent le discours et le texte. Cette navigation doit aller dans les deux sens : de la localité, l'occurrence, vers les structures et les éléments qu'elles connectent, d'une part et, d'autre part, de la structure, par exemple le plan du texte, vers ses parties constituantes. Ces parcours sont l'extension de notre pratique actuelle qui nous plonge du contexte au lexique, du lexique au contexte, une extension aussi des parcours des réseaux de co-occurrents et des réseaux lexicaux.

Ce premier type de considérations, justifiant notre proposition d'annotation structurelle, est complété par des considérations d'ordre documentaire. La *mise en connexion* n'est pas seulement *intratextuelle* : elle est aussi *intertextuelle*. Les textes font référence les uns aux autres, directement ou par le partage de mêmes paradigmes. Plus encore, l'analyse textuelle, en tant qu'elle-même pratique discursive, produit des textes sur des textes, des annotations sur des textes, y compris des textes d'annotation et d'analyse. Notre entreprise de modélisation doit donc aussi comporter une dimension documentaire permettant de mettre en relation les textes qui circulent dans l'espace public et autour desquels s'articule le discours social. Voilà pourquoi, du point de vue de son inscription concrète dans l'espace public, nous proposons que l'annotation analytique, commentaires, catégories ou graphes, prenne la forme de documents d'annotation XML respectant une syntaxe conforme aux recommandations du Text Encoding Initiative (TEI). Ces documents pourront ainsi s'intégrer plus aisément au modèle de dépôt de données adapté à la constitution de corpus de recherche (Daoust et al., 2008). Ces systèmes de dépôt de données, surtout connus pour la diffusion des publications scientifiques, peuvent être étendus aux résultats et procédures d'analyse au-delà de leur synthèse dans les articles scientifiques.

3. Documents d'annotation en TEI

3.1. Les propositions de Sacacomie

Un document d'annotation est une ressource électronique possédant un identifiant unique, au sens du W3C, et qui utilise des mécanismes de pointage permettant de faire référence à des parties d'un ou de plusieurs autres documents numériques aussi localisables par les mécanismes

standards du Web (URI et URL). On utilise le terme d'annotation dans son sens le plus large comprenant aussi le simple fait de commenter et de citer une ressource. On peut qualifier les documents d'annotation de *secondaires* par rapport aux documents annotés que l'on pourrait qualifier de *primaires*. Bien sûr, un document, considéré à une étape donnée comme *secondaire,* deviendra *primaire* par rapport à un autre document *secondaire* qui l'annoterait.

Le langage de balisage XML est maintenant l'approche privilégiée pour constituer des documents structurés ou semi-structurés en offrant une syntaxe unique et extensible selon des principes bien définis. La *Text Encoding Initiative* (TEI) est ce consortium qui se consacre depuis 1987 à formuler des propositions pour l'encodage des textes en format numérique pour la communauté des sciences humaines. Depuis leur version 3, les propositions de la TEI sont exprimées dans une syntaxe XML.

L'adoption des recommandations de la TEI par un grand nombre d'organismes dans le monde nous a incités, tout naturellement, à nous référer à ces recommandations pour proposer des formats XML-TEI pour l'échange de corpus et de ressources textuelles au sein des communautés qui gravitent autour des JADT. C'est ainsi que le réseau *ATONET* (2005) a proposé un sous-ensemble de balises TEI pour traduire, à des fins d'échange, les formats propriétaires utilisés par les logiciels d'analyse textuelle couramment employés au sein de la communauté de la recherche. C'est, ce que nous avons appelé les *propositions de Sacacomie* (Daoust and Marcoux, 2006), du nom du lieu où s'est tenu le séminaire présentant ces propositions.

Les *propositions de Sacacomie* comprennent un encodage dit *embarqué* (*embedded* en anglais) des annotations simples. Cela signifie que les annotations peuvent s'inscrire dans le document primaire selon la pratique de la majorité des logiciels considérés par le groupe de travail d'ATONET : Alceste (Reinert, 2002), Diatag-Astartex (Viprey, 2009), DTM (Lebart, 2005), Lexico (Salem et al., 2003) et SATO (Daoust, 2009). En fait, nous formulions à l'époque deux propositions : une *proposition de base* servant de commun dénominateur aux logiciels existants et une *proposition avancée* comprenant un découpage en mots marqué par la paire de balises <*w*> </*w*>. L'élément *w* est accompagné d'un attribut *xml:id* identifiant chacun des mots de manière unique. Cette proposition comprenait aussi le principe de document d'annotation externe utilisant les structures de traits (avec leur élément *fs « feature structure»*) pour annoter les formes lexicales et leurs occurrences. Notre proposition de format pour l'annotation structurelle s'appuie sur cette *proposition avancée de Sacacomie.* Elle reprend l'utilisation de l'élément *span* suggéré par la TEI pour référer, dans le document secondaire d'annotation, à un empan textuel dans le document primaire annoté.

Cet élément *span* est présenté dans le chapitre intitulé *Simple Analytic Mechanisms* du TEI P5: Guidelines (TEI Consortium 2007). Il y est décrit comme un des mécanismes simples de référence à des empans textuels utilisés à des fins analytiques. Il permet d'associer une annotation interprétative à un passage de texte référé par des pointeurs. Les <*span*> peuvent être coiffés d'un élément <*spanGrp*>, comme illustré dans l'exemple suivant.

```
<spanGrp resp="#Adam2005" type="ThèmeRhème" xml:base="http://monsite.org/doc-source.xml">
<span from="#w1" to "#w4" xml:id="Th1" ana="#thème"> Thème initial en début de phrase ( "Et un jour " ) </
span>
</spanGrp>
```

La balise <*spanGrp resp="#Adam2005" type="ThèmeRhème" xml:base= "http://monsite.org/doc-source.xml"*> permet de factoriser des attributs communs à un ensemble de <*span*> : *resp* renvoie à la description, généralement dans l'entête TEI, de la personne responsable de cette annotation, alors que *type* indique de quel type d'annotation il s'agit. L'attribut *xml:base*

contient l'URL du document analysé. Dans l'exemple, il s'agit du nom du document *doc-source.xml* sur *monsite.org*. On assume ici que ce document contient le texte à analyser découpé en mots identifiés par l'attribut *xml:id* des éléments <*w*>.

Le contenu de la balise <*span*> est utilisé pour délimiter un passage et expliquer la nature de l'annotation concernée. Les attributs *from* et *to* contiennent un pointeur sur le début et la fin du passage sur lequel porte l'annotation (l'attribut *to* est facultatif si le passage ne comporte qu'un élément). Dans l'exemple, *w1* et *w4* renvoient aux valeurs de l'attribut *xml:id* des éléments <*w*> dans le document primaire *doc1.xml*. Le <*span*> désigne donc de façon simple une étendue textuelle allant d'un mot à un autre, chacun des mots étant identifié par une étiquette unique dans le document référé ici par l'attribut *xml:base*. L'attribut *ana* pointe sur une interprétation de l'élément. Il est courant d'inscrire cette interprétation dans un élément <*interp*>. Les recommandations de la TEI indiquent que cet élément <*interp*> vise à résumer l'interprétation d'une annotation analytique. L'élément <*interp*> peut faire partie d'un <*interpGrp*> qui permet aussi de factoriser des attributs communs à un ensemble de balises <*interp*>. Ici, on fait appel à la combinaison des éléments <*span*> et <*interp*> pour distinguer le schéma général de l'analyse, avec la définition des concepts, de l'instanciation du concept sur un passage donné. La TEI signale qu'on pourrait aussi utiliser des structures de traits, plutôt que des éléments <*interp*>. Les structures de traits sont particulièrement appropriées lorsque l'analyse renvoie à des systèmes catégoriels. Donc, la TEI nous fournit tous les éléments et les attributs qu'il nous faut dans un ensemble bien documenté et diffusé dans la communauté des sciences humaines.

3.2. Première illustration : la relation thème-rhème

Dans les paragraphes qui suivent, nous présenterons un premier exemple de document TEI illustrant l'application de la *perspective fonctionnelle de la phrase* sur une courte phrase extraite d'Adam 2005 : 49. Voici la phrase et le schéma (schéma 8).

Schéma 8

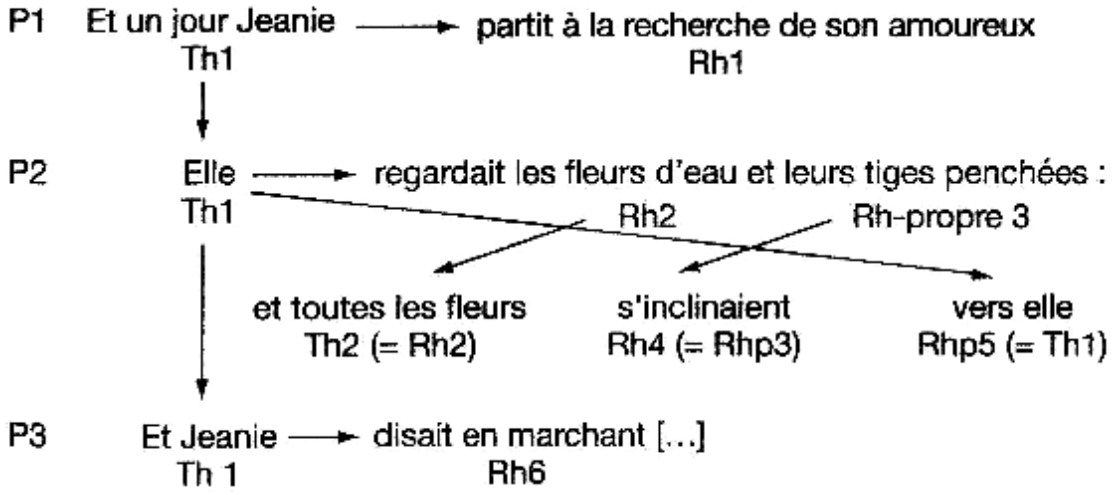

Figure 1 : Exemple de relation thème-rhème (schéma 8 extrait d' Adam 2005 : 49)

Et un jour Jeanie partit à la recherche de son amoureux. Elle regardait les fleurs d'eau et leurs tiges penchées : et toutes les fleurs s'inclinaient vers elle. Et Jeanie disait en marchant

Voici comment nous pourrions inscrire ce texte dans un document primaire XML-TEI conforme à la *proposition avancée de Sacacomie* (doc1.xml).

```
<?xml version=”1.0” encoding=”utf-8”?>
<TEI xmlns=”http://www.tei-c.org/ns/1.0”>
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title>Texte utilisé pour exemplifier une analyse fonctionnelle de type thème-rhème (Phébus, 1884,
2002 :429) : version électronique</title>
    <respStmt> <resp>mis en forme par</resp> <name>François Daoust</name> </respStmt>
  </titleStmt>
  <publicationStmt>
    <distributor>Université du Québec à Montréal, Centre ATO</publisher>
    <pubPlace>Québec, Canada</pubPlace> <date>2008-02-05</date>
  </publicationStmt>
  <notesStmt>
    <note>Des annotations analytiques sur le texte figurent dans des fichiers séparés.</note>
  </notesStmt>
  <sourceDesc> <bibl> Adam, Jean-Michel. La linguistique textuelle, Introduction à l'analyse textuelle des
discours. Page 49. Armand Colin, Paris 2005, ISBN 2-200-26752-5.</bibl> </sourceDesc>
  </fileDesc>
  <profileDesc> <langUsage> <language ident=”fr”>Français</language> </langUsage> </profileDesc>
  <encodingDesc>
   <refsDecl>
    <p> Le texte est découpé en pages (élément vide pb), ligne (élément vide lb) et mots (élément w).</p>
   </refsDecl>
  </encodingDesc>
 </teiHeader>
 <text>
  <body>
<pb n=”49”/>
<p> <lb/><w xml:id=”w1”>Et</w> <w xml:id=”w2”>un</w> <w xml:id=”w3”>jour</w> <w
xml:id=”w4”>Jeanie</w> <w xml:id=”w5”>partit</w> <w xml:id=”w6”>à</w> <w xml:id=”w7”>la</
w> <w xml:id=”w8”>recherche</w> <w xml:id=”w9”>de</w><w xml:id=”w10”>son</w><w
xml:id=”w11”>amoureux</w><w xml:id=”w12”>.</w>
<w xml:id=”w13”>Elle</w> <w xml:id=”w14”>regardait</w> <w xml:id=”w15”>les</
w> <w xml:id=”w16”>fleurs</w>  <lb/><w xml:id=”w17”>d'</w><w xml:id=”w18”>eau</
w> <w xml:id=”w19”>et</w> <w xml:id=”w20”>leurs</w> <w xml:id=”w21”>tiges</w> <w
xml:id=”w22>penchées</w> <w xml:id=”w23”>:</w> <w xml:id=”w24”>et</w> <w xml:id=”w25”>toutes</
w> <w xml:id=”w26”>les</w> <w xml:id=”w27”>fleurs</w> <w xml:id=”w28”>s'</w><w
xml:id=”w29”>inclinaient</w> <w xml:id=”w30”>vers</w> <w xml:id=”w31”>elle</w><w
xml:id=”w32”>.</w>
<w xml:id=”w33”>Et</w> <w xml:id=”w34”>Jeannie</w> <lb/><w xml:id=”w35”>disait</w> <w
xml:id=”w36”>en</w> <w xml:id=”w37”>marchant</w> <!-- etc. → </p>  </body>
  </text>
</TEI>
```

doc1.xml

Et voici maintenant un document d'annotation externe (ana1.xml) décrivant les relations thèmes-rhèmes présentées dans le schéma 8 de Jean-Michel Adam.

```
<?xml version="1.0" encoding="utf-8"?>
<TEI>
  <teiHeader> <!-- etc. → </teiHeader>
  <text><body>
    <div type="Analyse" subtype="ThèmeRhème" xml:id="ana1">
<interp xml:id="Thème">Le thème est l'énoncé qui se pose comme connu</interp>
<interp xml:id="Rhème">Le rhème est un énoncé qui ajoute de l'information sur un énoncé thème</interp>
<interp xml:id="ThèmeConstant">Le thème constant correspond à une progression thématique dans lequel un même thème est repris dans une suite de relations thèmes-rhèmes</interp>
<interp xml:id="ThématisationLinéaire">La thématisation linéaire correspond à une progression thématique dans laquelle un rhème est repris à titre de thème dans la succession des énoncés.</interp>

<!-- Les relations Thèmes-Rhèmes  -->
<ab xml:id="TR1" type="ThèmeRhème" xml:base="doc1.xml">
  <span ana="#Thème" xml:id="T1-4" from="#w1" to="#w4" n="Th1">
     Et un jour Jeanie (thème initial en début de phrase )
  </span>
  <span ana="#Rhème" xml:id="R5-12" from="#w5" to="#w12" n="Rh1">
     partit à la recherche de son amoureux.
  </span>
</ab>
<ab xml:id="TR2" type="ThèmeRhème" xml:base="doc1.xml">
  <span ana="#Thème" xml:id="T13-13" from="#w13" to="#w13"  n="Th1">Elle</span>
  <span ana="#Rhème" xml:id="R14-18" from="#w14" to="#w18" n="Rh2">regardait les fleurs d'eau</span>
  <span ana="#Rhème" xml:id="R19-23" from="#w19" to="#w23" n="Rhp3">et leurs tiges penchées:</span>
</ab>
<ab xml:id="TR3" type="ThèmeRhème" xml:base="doc1.xml">
  <span ana="#Thème" xml:id="T24-27" from="#w24" to="#w27"  n="Th2">et toutes les fleurs (=Rh2)</span>
  <span ana="#Rhème" xml:id="R28-29" from="#w28" to="#w29" n="Rh4">s'inclinaient (=Rhp3)</span>
  <span ana="#Rhème" xml:id="R30-32" from="#w30" to="#w32" n="Rhp5">vers elle. (=Th1)</span>
</ab>
<ab xml:id="TR4" type="ThèmeRhème" xml:base="doc1.xml">
  <span ana="#Thème" xml:id="T33-34" from="#w33" to="#w34"  n="Th1">Et Jeannie</span>
  <span ana="#Rhème" xml:id="R35-37" from="#w35" to="#w37" n="Rh6">disait en marchant</span>
 </ab>

<!-- Les progressions thématiques. -->
<ab xml:id="PT1" type="ProgressionThématique" ana="#ThèmeConstant">
  <span from="#T1-4" n="Th1">« Et un jour Jeanie " : thème initial en début de phrase </span>
  <span from="#T13-13" n="Th1">« Elle " : anaphore pronominale</span>
  <span from="#T33-34" n="Th1">« Et Jeanie " : reprise </span>
</ab>
<ab xml:id="PT2" type="ProgressionThématique" ana="#ThématisationLinéaire">
  <span from="#R14-18" n="Rh2">« regardait les fleurs d'eau " </span>
  <span from="#T24-27" n="Th2">« et toutes les fleurs " </span>
</ab>
<ab xml:id="PT3" type="ProgressionThématique" ana="#ThématisationLinéaire">
  <span from="#R30-32" n="Rhp5">« vers elle " </span>
  <span from="#T13-13" n="Th1">« elle " (chiasme qui rhématise  le pronom anaphorique de PT1)</span>
</ab>
    </div>
  </body> </text>
</TEI>
```

ana1.xml

Après l'entête TEI, le corps du document comprend un élément <*div*> (division) avec un attribut *(subtype)* qui indique le type d'analyse effectué et un identifiant pour ce bloc d'analyse dans l'attribut *xml:id*. La valeur *ThèmeRhème* de l'attribut *type* dans <*ab*> indique la nature de la relation décrite dans le bloc. La valeur de l'attribut *xml:id* identifie chacune des relations et l'attribut *xml:base* indique sur quel document portent les références de la relation. On retrouve ensuite des éléments *interp*, avec leur identifiant dans l'attribut *xml:id*, qui contiennent des explications sur les catégories de l'analyse.

On retrouve ensuite des blocs (élément *ab* pour *arbitrary bloc*) qui décrivent les diverses relations de type ThèmeRhème. Dans les <*span*> qui définissent les empans référés par l'analyse, on utilise l'attribut *ana* pour pointer vers la catégorie analytique appliquée à l'empan, ici un texte libre dans un élément <*interp*>. Le contenu textuel des *span* n'est là qu'à titre informatif pour faciliter la lecture sans retourner au texte primaire.

Les relations thèmes-rhèmes se complètent par des relations de progression thématique reliant les thèmes entre eux. La structure de progression linéaire, par exemple, indique qu'un élément rhématisé est repris à titre de thème dans une autre relation. La progression thématique réutilise les segments déjà décrits, mais dans des constructions différentes. Ainsi, dans l'exemple, on retrouve l'utilisation d'éléments <*ab*> de type *progression_thématique*. L'attribut *ana* précise le type de progression impliquée. Les <*span*> reprennent les énoncés impliqués dans la structure. Le contenu des éléments permet d'apporter des commentaires explicatifs destinées au lecteur humain. L'attribut *n* reprend simplement les étiquettes symboliques utilisées par Adam.

3.3. Deuxième illustration : la structure compositionnelle d'un texte

La relation thème-rhème, même si elle peut dépasser la frontière de la phrase, couvre un empan textuel relativement restreint. À l'opposé, la structure compositionnelle d'un texte recouvre l'ensemble du texte. Adam nous en donne un exemple sur un court récit de Jorge Luis Borges, El Hacedor traduit par J.-M. Adam (2005 : 203-204).

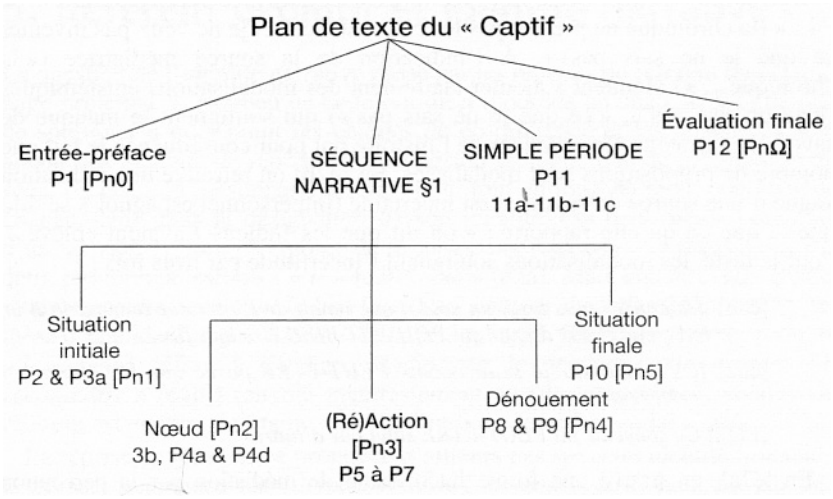

Figure 2 : Plan de texte du « Captif » (extrait d'Adam 2005 : 203-204)

Pour représenter cette analyse d'Adam en XML-TEI, nous avons d'abord balisé le texte source selon le format *Sacacomie.* En plus du découpage en mots, le texte contient un découpage en énoncés et en phrases. Le document d'annotation externe suivant présente ces découpages et la

structure compositionnelle du texte précédée d'éléments <*interp*> qui décrivent les catégories de l'analyse.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader> <!-- etc → </teiHeader>
 <text> <body>
  <!-- Définition des catégories interprétatives -->
  <div type="Analyse" subtype="SC" xml:id="ana1">
   <interpGrp type="Unités_discursives">
    <interp xml:id="Énoncé">On considèrera comme énoncé...</interp>
    <interp xml:id="Phrase">On entendra par phrase typographique...</interp>
   </interpGrp>
   <interpGrp type="SC">
    <interp xml:id="plan_de_texte">Le plan du texte fait partie de la structure compositionnelle qui organise la cohésion d'une suite linéaire de séquences (Adam2005:chapitre 6).</interp>
     <interp xml:id="séquence">Les séquences sont des unités textuelles complexes, composées d'un nombre limité de paquets de propositions-énoncés. Elles constituent des réseaux relationnels hiérarchiques formant des entités relativement autonomes présentant des agencements dits narratifs, argumentatif, explicatif, dialogal, etc. (Adam2005:chapitre 5). </interp> <!-- etc -->
   </interpGrp>
<!-- Segmentation du texte analysé en énoncés -->
   <spanGrp xml:id="Seg3" type="Segmentation" ana="#Énoncé" xml:base="borges_adam.xml">
     <span from="#w2" to="#w13" xml:id="é1">À Junín ou à Tapalqué, on raconte l'histoire suivante.</span> <span from="#w14" to="#w23" xml:id="é2a">Un enfant disparut après un raid d'Indiens ;</span> <span from="#w24" to="#w33" xml:id="é2b">on dit que les Indiens l'avaient enlevé.</span> <!-- etc -->
   </spanGrp>
<!-- Segmentation du texte analysé en phrases -->
   <spanGrp xml:id="Seg4" type="Segmentation" ana="#Phrase">
    <span from="#é1" xml:id="P1"/> <span from="#é2a" to="#é2b" xml:id="P2"/> <!-- etc →
   </spanGrp>
<!-- Composants de l'analyse compositionnelle -->
   <div type="SC" ana="#SCsequence_narrative" xml:id="séquence_narrative_1"> <!-- etc → </div>
   <div type="SC" ana="#SCsequence_narrative" xml:id="entrée-préface"> <!-- etc --> </div>
   <div type="SC" ana="#SCpériode_argumentative" xml:id="P11_argumentative">
    <span type="SC" from="#é11a" xml:id="é11a-proposition_p"> premier argument </span>
     <span type="SC" from="#é11b" xml:id="é11b-proposition_q">second argument</span>
   <span type="SC" from="#é11c" xml:id="é11c-conclusion_non_c">renversement de la conclusion implicite du retour définitif à la maison</span>
   </div>
   <div type="SC" ana="#SCpériode_narrative" xml:id="P11_narrative">
    <span type="SC" from="#é11a" xml:id="é11a-Pn1">Situation initiale Pn1</span>
    <span type="SC" from="#é11b" xml:id="é11b-Pn2">Nœud Pn2</span>
    <span type="SC" from="#é11c" xml:id="é11c-Pn4">Dénouement Pn4</span>
   </div>
   <div type="SC" ana="#SCpériode" xml:id="simple_période_P11">
    <alt mode="incl" targets="#P11_argumentative #P11_narrative" weights="0.5 0.5"/>
   </div>
   <div type="SC" ana="#SCpériode" xml:id="évaluation_finale">
    <span type="SC" from="#P12" xml:id="PnΩ">Évaluation finale. «Cette prose périodique dominée par le rythme contribue au glissement de genre du récit factuel au récit poétique." (Adam 2005: 211)</span>
   </div>
<!-- Bloc supérieur : plan du texte -->
   <div type="SC" ana="#SCplan_de_texte" xml:id="plan_de_texte_du_Captif">
    <ab>
     <ptr target="#entrée-préface"/>
```

```
        <ptr target="#séquence_ narrative_1"/>
         <ptr target="#simple_période_P11"/>
         <ptr target="#évaluation_finale"/>
       </ab>
     </div>
   </div>
  </body></text>
</TEI>
```

Dans le bloc *Segmentation du texte analysé en énoncés* du document d'annotation, on a mis le texte référé par les *span* à titre explicatif puisque la référence aux empans textuels dans le document annoté suffit à recomposer le texte. Ces divers segments phrastiques ou propositionnels sont organisés à des fins d'analyse en plusieurs regroupements périodiques et un regroupement séquentiel.

La composition structurelle emprunte donc ici la forme classique de l'emboîtement d'éléments TEI <*div*> (division) portant l'attribut *type=SC.* On utilise *les divisions* comme on le ferait pour décrire la structure formelle d'un texte sauf que, cette fois-ci, le contenu textuel des divisions est constitué de références à des segments dont les pointeurs, une fois évalués, conduiront finalement à des empans textuels dans le document analysé. Il s'agit en quelque sorte de divisions à portée analytique à l'intérieur d'un document d'analyse portant sur un texte, objet de l'analyse, qui est contenu dans une ressource externe. Les valeurs de l'attribut *ana* renvoient à des explications sur l'interprétation de chaque structure compositionnelle (éléments *interp).*

Dans l'exemple, on trouve deux structures pleinement exposées. On a *P11_argumentative* avec ses trois *span* correspondant à deux arguments et à une conclusion. Et on a *P11_narrative* contenant trois empans textuels correspondant à la situation, au nœud et au dénouement de la période narrative. En fait, ces deux structures sont deux points de vue sur la même portion du texte. Aussi, la division suivante (*simple_période_P11*) indique (élément *alt*) que ces interprétations sont possibles en même temps (*mode="incl"*) à part égale (*weights="0.5 0.5"*). Les deux analyses ne sont pas directement incluses dans l'élément *alt*, mais elles sont référés par des pointeurs sur les éléments *div* précédemment décrits.

Finalement, la division *plan_de_texte_du_Captif* rassemble (via l'élément *ab* pour *arbitrary bloc*) sous forme de pointeurs (élément *Ptr*) tous les épisodes par des références aux divisions d'analyse déjà décrites.

Cette construction du plan du texte par *modules* est une formalisation directe d'un processus d'analyse qui relève d'un va et vient entre la reconnaissance d'éléments macrostructurels, leur décomposition en structures plus fines jusqu'aux propositions-énoncés, et leur rassemblement dans un plan de texte englobant.

4. Conclusion et perspectives

Le recours aux recommandations de la TEI pour réaliser des documents d'annotation en général, et d'annotation structurelle en particulier, nous semble une voie prometteuse pour la diffusion et l'interopérabilité des traitements sur corpus. Le partage des mêmes formalismes pour l'édition électronique des corpus et pour la production de documents d'analyse sur les corpus traduit bien la réalité discursive de « textes sur les textes », qui se répondent et s'entrecroisent.

Certes, la représentation XML d'un document d'annotation structurelle, même si elle est directement lisible par l'humain, n'est pas la représentation privilégiée du point de vue

ergonomique. Aussi, nous pouvons appliquer une feuille de style XSLT qui transforme cette représentation en une autre représentation XML qui traduit le formalisme décrit en graphes constitués de nœuds et d'arcs entre les nœuds. Cette représentation peut alimenter des librairies graphiques qui traceront le graphe à la manière des figures qui illustrent les exemples de Jean-Michel Adam. Ainsi, Serge Fleury a déjà intégré dans son logiciel *Le Trameur* (Fleury, 2009) un module capable de produire ces représentations graphiques. Plus encore, il offre déjà des fonctions pour ajouter des nœuds et des arcs entraînant des modifications équivalentes dans la structure XML sous-jacente.

L'interface graphique pour l'affichage et la construction des annotations structurelles n'est pas le seule approche possible. On pourrait, par exemple, souhaiter disposer d'*assistants* facilitant l'instanciation des schémas de structures. On pourrait associer à ces schémas des contraintes de divers ordres (position relative dans le texte, valeurs dans une structure de traits, etc.) permettant de valider les structures. Au-delà de ces contraintes d'intégrité, il s'agirait de faciliter le dépistage de segments susceptibles d'occuper des positions manquantes dans des structures partiellement construites. Par exemple, la reconnaissance de marqueurs de liaison pourrait déclencher l'instanciation de structures dont les arguments resteraient à repérer. Ainsi, une occurrence de l'adverbe *premièrement* suggère une structure énumérative ou argumentative dont les termes devraient aussi être introduits par des marqueurs de relation d'un certain type porté par un trait lexical. Cela nous ramène à l'idée des *marques instructionnelles* dont parle Adam. Le genre du texte analysé induit aussi des attentes de lecture et des stratégies de repérage des composants de la structure attendue. Donc, les instructions de lecture peuvent provenir de son contenu linguistique interne, mais aussi du contexte communicationnel dans lequel s'inscrit le texte.

L'annotation, en particulier dans sa composante structurelle, est donc un processus interactif et itératif qui exigera des stratégies d'optimisation informatique permettant un temps de réponse acceptable du point de vue de l'interaction de l'analyste avec son corpus. Il faudra aussi trouver le moyen de faciliter l'usage des langages d'interrogation et de mises à jour comme *XQuery* et *Xquery Update Facility*. Des formalismes comme *XPath* et *XSLT* sont aussi des outils puissants mais leur maîtrise directe par le lecteur-analyste pose des problèmes. On fait donc face à de nombreux défis pour rendre accessible l'annotation structurelle : des défis sur les formalismes des données et des langages de requête, sur les stratégies d'implantation informatique, sur l'ergonomique et sur l'apprentissage. Cependant, l'intensité des recherches dans le monde XML est telle que les ressources disponibles pour relever ces défis se développent rapidement.

Références

Adam J.-M. (2005). *La linguistique textuelle, Introduction à l'analyse textuelle des discours*. Paris : Armand Colin.

ATONET (2005). *Réseau pour l'échange de ressources et de méthodologies en analyse de texte assistée par ordinateur* (ATONET) : http://www.atonet.net.

Bakhtine M. (1984). *Esthétique de la création verbale*. Paris : Gallimard.

Charolles M. (1993). Les plans d'organisation du discours et leur interaction, in Moirand, S., Bouacha, A.A., Beacco, J.-C. and Collinot, A., editors, *Parcours linguistiques de discours spécialisés*, Berne : Peter Lang, pp. 301-314.

Daoust F. (2009). Système d'analyse de texte par ordinateur, SATO, Manuel de référence, version 4.3. Centre d'analyse de texte par ordinateur, UQAM, 2007; modifié en 2009. http://www.ling.uqam.ca/sato/satoman-fr.html.

Daoust F., Duchastel J., Marcoux Y. and Rizkallah E. (2008). JADT-2008. Pour un modèle de dépôt de données adapté à la constitution de corpus de recherche. In *Actes des JADT-2008*, vol. 1, pp. 355-367, Presses universitaires de Lyon, 2008. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/daoust-duchastel-marcoux-rizkallah.pdf.

Daoust F. and Marcoux Y. (2006). Logiciels d'analyse textuelle : vers un format XML-TEI pour l'échange de corpus annotés. In *Les Cahiers de la MSH Ledoux no. 3, Actes des JADT-2006*, vol. 1, pp. 327-340, Presses universitaires de Franche-Comté, 2006. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/PDF/029.pdf.

Fleury S. (2009). *Le métier textométrique* (Trameur). Centre de textométrie – CLA^{2}T, U. Paris 3 Sorbonne nouvelle, http://tal.univ-paris3.fr/trameur/.

Habert B. (1998). Des mots complexes possibles aux mots complexes existants : l'apport des corpus, Mémoire présenté pour l'obtention d'une habilitation à diriger des recherches. Document de synthèse, Université Lille III - Charles de Gaulle http://www.limsi.fr/Individu/habert/Publications/Fichiers/hdr/node4.html.

Lebart L. (2005). *Data and Text Mining.* École nationale supérieure de télécommunications, Paris, http://www.enst.fr/egsh/lebart/.

Reinert M. (2002). *Alceste, Manuel de référence*. Université de Saint-Quentin-en-Yvelines, CNRS.

Salem A., Lamalle C., Martinez W., Fleury S., Fracchiolla B., Kuncova A. and Maisondieu A. (2003). *Lexico3 – Outils de statistique textuelle. Manuel d'utilisation*. Syled-CLA2T, Université de la Sorbonne nouvelle – Paris 3 : http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW.

TEI Consortium (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium, eds. http://www.tei-c.org/Guidelines/P5/.

Viprey J-M. (2009). *DiaTag–Astartex.* Université de Franche-Comté. http://laseldi.univ-fcomte.fr/document/viprey/page_JMV.htm.

W3C (2009a). *XQuery Update Facility* 1.0. W3C W3C Candidate Recommendation 09 June 2009. http://www.w3.org/TR/2007/REC-xquery-20070123/.

W3C (2007a). *XML Path Language (XPath) 2.0.* W3C Recommendation 2007. http://www.w3.org/TR/2007/REC-xpath20-20070123/.

W3C (2007b). *XQuery 1.0: An XML Query Language.* W3C Recommendation 2007. http://www.w3.org/TR/2007/REC-xquery-20070123/.

W3C (2007c). *XSL Transformations (XSLT) Version 2.0.* W3C Recommendation 2007. http://www.w3.org/TR/2007/REC-xslt20-20070123/.

Weinrich H. (1964/1973). *Le temps.* Paris : Seuil [cité par Adam, 2005].