Une typologie multi-dimensionnelle des structures énumératives pour l'identification des relations termino-ontologiques

Jean-Philippe Fauconnier, Mouna Kamel, Bernard Rothenburger

Une typologie multi-dimensionnelle des structures énumératives pour l'identification des relations termino-ontologiques

Jean-Philippe Fauconnier[1] **Mouna Kamel**[1] **Bernard Rothenburger**[1]

[1] Institut de Recherche en Informatique de Toulouse (IRIT)
Université Paul Sabatier, 118 Route de Narbonne, 31060 Toulouse Cedex 5
{prénom}.{nom}@irit.fr

Résumé

Ce travail s'inscrit dans le cadre de la construction de ressources termino-ontologiques. Il vise à améliorer l'extraction des relations sémantiques en exploitant les structures énumératives contenues dans les textes. Nous proposons ici une typologie multi-dimensionnelle de ces structures énumératives, selon les axes visuel, rhétorique, intentionnel et sémantique. Cette typologie intervient dans le cadre d'une campagne d'annotation outillée par LARAt (Logiciel d'Acquisition de Relations par l'Annotation de textes), pour l'identification de relations par apprentissage supervisé.

1 Introduction

La structure énumérative (dorénavant appelée SE) est une structure textuelle ayant la propriété d'exprimer des connaissances hiérarchiques au travers de différents composants. Elle présente, au sein d'un même objet textuel, un thème énumératif, dit *énumérathème*, justifiant la réunion de plusieurs éléments en fonction d'une identité de statut (Ho-Dac et al., 2010). Sur le plan sémantique elle forme un tout. Sur le plan de la mise en forme, elle peut être exprimée selon différents modes, allant d'une forme linéaire discursive à une forme visuelle usant de dispositifs typo-dispositionnels. Ces propriétés autorisent son apparition dans tout type de texte, lui permettant par là même de rendre compte de connaissances de nature différente.

Elle a ainsi fait l'objet de nombreuses études au cours desquelles différentes typologies ont pu être proposées. Les SE linéaires ont été essentiellement analysées dans le cadre de l'analyse du discours. Elles ont d'abord donné lieu à des typologies comme celle de (Vergez-Couret et al., 2008) où les SE à un temps ont été opposées aux SE à deux temps, ou encore comme celle de (Ho-Dac et al., 2010) où les SE ont été classifiées selon leur niveau de granularité (SE dont les items sont des titres, SE en tant que listes formatées, SE multi-paragraphiques sans marque visuelle, SE intra-paragraphiques). Les SE usant de dispositifs typo-dispositionnels, dites verticales, ont quant à elles été notamment analysées dans le cadre de la génération de texte. Hovy et Arens (1991) distinguent les listes d'items (ensemble de composants de même niveau), des listes énumérées (pour lesquelles l'ordre des composants est pris en compte), alors que Luc (2001) propose une typologie qui oppose les SE parallèles aux SE non parallèles. Cette dernière typologie est basée sur la composition du modèle rhétorique de la RST[1] (Mann and Thompson, 1988) et du MAT[2] de Virbel (1989).

À notre connaissance, les SE n'ont pas été exploitées pour l'extraction de relations sémantiques à partir de textes. Or ces SE sont très fréquentes dans les textes scientifiques ou encyclopédiques qui sont justement appropriés pour la construction de ressources sémantiques. Les méthodes classiques d'extraction des relations sont le plus souvent limitées à l'identification de relations binaires intra-phrastiques, après analyse du texte rédigé par des patrons lexico-

[1] RST : Rhetorical Structure Theory
[2] MAT : Modèle d'Architecture Textuelle

syntaxiques (Hearst, 1992; Montiel-Ponsoda and de Cea, 2011; Aussenac-Gilles and Jacques, 2008), des techniques de clusterisation ou des algorithmes d'apprentissage automatique (essentiellement non supervisé) (Buitelaar et al., 2005; Poelmans et al., 2010). L'exploitation des SE apparait alors comme un moyen d'élargir les méthodes classiques d'extraction de relations pour la construction ou l'enrichissement de ressources sémantiques telles que les ontologies, les Ressources Termino-Ontologiques (RTO), les thesaurus, etc.

Cet article propose une typologie multi-dimensionnelle qui permettra de cibler puis d'exploiter automatiquement les SE porteuses de relations termino-ontologiques. Cette typologie caractérise les SE selon les axes visuel et rhétorique à l'instar de (Luc, 2001), mais également selon les axes intentionnel et sémantique. C'est cette typologie que nous présentons en section 3, après avoir rappelé en section 2 quelques définitions et propriétés des SE. Vu l'inadéquation des outils classiques d'extraction de relations pour ce genre de structure textuelle, nous envisageons une approche alternative, à base d'apprentissage supervisé, nécessitant une campagne d'annotation basée sur cette typologie. La section 4 montre comment cette typologie intervient dans le cadre du processus d'annotation, et décrit sommairement l'outil d'annotation développé pour ces besoins. Nous concluons et présentons nos perspectives en section 5.

2 SE : définitions et propriétés

Comme indiqué précédemment, l'acte d'énumération consiste à énoncer les éléments successifs d'un même champ conceptuel, ces éléments entretenant un lien hiérarchique direct ou indirect avec un concept classifieur. La forme générale d'une SE est alors caractérisée par la présence d'une *amorce* (phrase contenant l'énumérathème et introduisant l'énumération), d'une *énumération* composée d'au moins deux *items* (appartenant au même champ conceptuel), et éventuellement d'une *clôture* (ou conclusion).

D'un point de vue visuel, la SE a la propriété de pouvoir être formulée de diverses façons. Elle peut être énoncée discursivement en dehors de toute MFM, au sein de la même phrase ou à travers plusieurs phrases n'appartenant pas nécessairement au même paragraphe. Elle peut également être mise en évidence par l'usage de marqueurs typographiques et/ou dispositionnels, marqueurs qui pallient alors les marqueurs lexicaux. Ces marqueurs sont de l'ordre de la métalangue (Harris, 1976; Porhiel, 2007) et permettent alors d'organiser des segments de texte successifs non forcément contigus.

Différentes définitions de la SE existent, dont celle de Pascual pour qui "énumérer, c'est conférer une égalité d'importance à un ensemble d'objets, et ensuite c'est ordonner ces objets selon des critères variés" (Pascual, 1991). Ces objets sont considérés comme visuellement et fonctionnellement équivalents. On parle alors de SE parallèles.

D'un point de vue rhétorique, l'analyse des SE montre qu'il existe des relations de discours entre les différents composants. La définition de Pascual citée ci-dessus correspond au cas où ces relations montrent une égalité d'importance entre les items. Or des études de corpus ont montré que les SE ne présentent pas toutes cette équivalence visuelle et fonctionnelle entre items (Luc, 2001).

Dans un souci de généralisation, nous préférons la définition proposée par (Virbel, 1999) qui nous semble mieux prendre en compte à la fois les phénomènes architecturaux du texte et l'intention de l'auteur : "l'acte textuel consiste à transposer textuellement la coénumérabilité des entités recensées par la coénumarabilité des segments linguistiques qui les décrivent, ceux-ci devenant par le fait les entités constitutives de l'énumération (les items)."

D'un point de vue intentionnel, à l'image des textes qui peuvent être de différents types (narratifs, procéduraux, descriptifs, etc.), les SE reflètent l'intention de l'auteur. Nous proposons de reprendre cette typologie des textes pour caractériser l'intention de l'auteur lorsqu'il rédige une SE.

Enfin, d'un point de vue sémantique, les SE peuvent exprimer des connaissances de nature différente. Ces connaissances peuvent décrire de

façon consensuelle ou conjoncturelle le monde réel ou imaginaire, la langue, les émotions, les sentiments, les opinions, etc.

3 Typologie de la SE

La typologie que nous proposons est basée sur les différentes propriétés décrites ci-dessus. Elle s'appuie sur les dimensions visuelle, rhétorique, intentionnelle et sémantique, l'objectif étant à terme de repérer et d'exploiter les SE paradigmatiques bénéficiant de mise en forme et véhiculant des connaissances propices à la construction de ressources sémantiques.

Les différentes caractéristiques observées au sein de chacune des dimensions sont illustrées par des exemples extraits du corpus de Virbel (1999) et d'un corpus composé de pages Wikipédia, ce deuxième corpus ayant été élaboré dans le but d'enrichir l'ontologie OntoTopo construite lors du projet GEONTO[3] (Kamel and Rothenburger, 2011).

3.1 Typologie selon l'axe visuel

Les types définis dans cet axe ont pour but d'aider au repérage des SE. Nous distinguons la **SE horizontale** qui peut bénéficier ou non de mise en forme typographique, de la **SE verticale** qui bénéficie de mise en forme typographique et dispositionnelle.

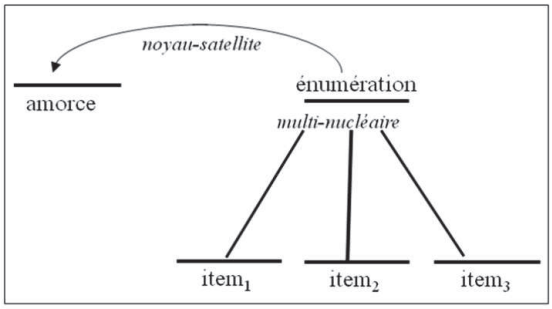
La **SE horizontale** s'inscrit dans la linéarité du texte et ne fait pas usage du "dispositionnel". Elle est caractérisée soit par des MIL[4] comme "premièrement", "deuxièmement", "d'abord", "ensuite", etc. qui permettent d'introduire les items (fig. 3.a), soit par des marqueurs lexicaux comme "tels que", "comme", etc. qui permettent d'introduire l'énumération (fig. 3.b). Mais elle peut aussi faire usage de marqueurs typographiques pour délimiter l'énumération, comme les parenthèses dans (fig. 3.c).

La **SE verticale** présente des discontinuités par rapport à la linéarité du texte. Des marqueurs typo-dispositionnels sont alors utilisés pour organiser, subdiviser et hiérarchiser les différents composants de la SE, comme le montre (fig. 3.d). Les items apparaissent en retrait par rapport à l'amorce, les items sont introduits par des puces, des tirets, etc.

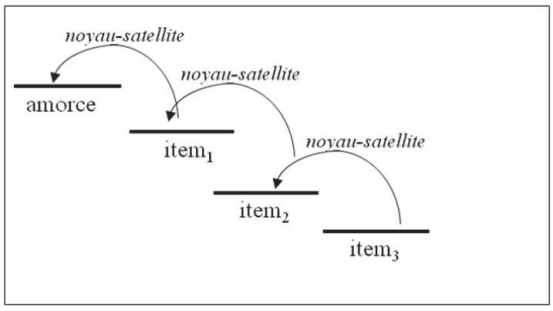
SE verticales et horizontales peuvent être combinées et imbriquées au sein d'une même SE. C'est le cas lorsqu'un item décrit lui-même une SE, avec ou sans mise en forme typo-dispositionnelle (fig. 3.e).

3.2 Typologie selon l'axe rhétorique

À ce niveau nous prenons en compte la nature des relations du discours qui relient les différents composants de la SE. Les relations entre items peuvent être de type noyau-satellite ou multi-nucléaire, selon la RST (Mann and Thompson, 1988). Une relation noyau-satellite relie une unité du discours plus saillante à une unité du discours qui supporte l'information d'arrière-plan, alors qu'une relation multi-nucléaire relie des unités du discours de même importance. Les SE, dont les items montrent une égalité d'importance, suscitent pour nous un intérêt particulier, car leur traduction en structures hiérarchiques est assez immédiate.



(a) structure rhétorique de la SE paradigmatique



(b) structure rhétorique de la SE syntagmatique

Figure 1: Représentations rhétoriques des SE paradigmatique et syntagmatique selon la RST.

Nous distinguons alors les **SE paradigmatiques**, les **SE syntagmatiques**, les **SE hybrides** et les **SE bivalentes**, reprenant ainsi en partie la terminologie utilisée par Luc (2001).

[3] ANR-07-MDCO-005, `http://geonto.lri.fr/`

[4] MIL : Marqueurs d'Intégration Linéaire

La **SE paradigmatique** est composée d'items indépendants dans un contexte donné. Elle porte alors une relation rhétorique multi-nucléaire entre les items successifs, chacun des items étant lié à l'amorce par une même relation de type noyau-satellite (fig. 1.a). Les exemples (a), (b), (c), entre autres, de la fig. 3 sont des cas de SE paradigmatiques. À l'opposé, la **SE syntagmatique** est composée d'items qui n'ont pas la même importance, et qui ne sont donc pas indépendants. La SE syntagmatique porte alors une relation rhétorique noyau-satellite entre items successifs (fig. 1.b). Le cas (fig. 3.f) en est un exemple.

Lorsqu'une SE porte une relation rhétorique noyau-satellite entre au moins deux items et une relation rhétorique multi-nucléaire entre au moins deux items, elle est qualifiée d'**hybride**. Enfin, les caractères paradigmatique et syntagmatique peuvent coexister au sein de la même SE, et dans ce cas la SE est dite **bivalente** (fig. 3.g).

3.3 Typologie selon l'axe intentionnel

À ce niveau nous prenons en compte l'intention de communication de l'auteur. Nous avons repris la typologie des textes pour l'adapter aux SE, en différenciant les **SE descriptives**, les **SE narratives**, les **SE prescriptives**, les **SE procédurales**, les **SE explicatives**, et les **SE argumentatives**. Ces types se sont révélés être les plus fréquents dans nos corpus. L'objectif est de caractériser les types de SE propices à la construction de RTO, pour ensuite proposer un modèle de représentation des connaissances adapté.

La **SE descriptive** décrit une entité qui peut être un objet du monde animé ou pas, artificiel ou naturel (fig. 3.a, fig. 3.b, fig. 3.c), alors que la **SE narrative** articule une succession d'actions ou d'événements, réels ou imaginaires (fig. 3.j). Les notions de conseil, d'indication, d'injonction peuvent être intégrées à ces types de SE. Dans ce cas la SE est dite **prescriptive** (fig. 3.i). De plus, lorsque ces conseils, indications, injonctions sont énoncés selon une volonté d'ordonnancer (comme dans les modes d'emploi, les notices explicatives, les guides d'utilisation, les manuels, les recettes de cuisine, etc.), pour atteindre un but donné, la SE est dite **procédurale** (fig. 3.h).

Enfin, la **SE explicative** répond en général à un questionnement de type "comment ?", "pourquoi?", "dans quelles circonstances?" etc. (fig. 3.f). Si des arguments sont avancés dans le but de défendre une opinion, dans le but de convaincre, la SE est dite **argumentative** (fig. 3.k).

En ce qui concerne cet axe, une même SE pourra posséder plusieurs traits intentionnels. La hiérarchie présentée en (fig. 2) décrit les combinaisons de types intentionnels les plus fréquentes.

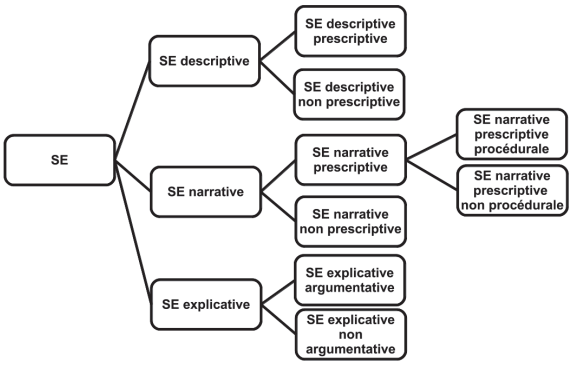


Figure 2: Combinaisons possibles des traits intentionnels au sein d'une même SE

Il existe cependant des SE pour lesquelles aucune des catégories de l'axe intentionnel précitées n'a pu être identifiée. Pour les catégoriser, nous avons défini le type **SE intentionnelle autre**.

3.4 Typologie selon l'axe sémantique

À ce niveau nous rendons compte de la dimension référentielle des SE, conformément à notre objectif de construction de ressources termino-ontologiques. Nous avons divisé les SE en trois catégories : **SE à visée ontologique** concerne des connaissances du monde (fig. 3.d et fig. 3.g), **SE métalinguistique** concerne la langue (fig. 3.l et fig. 3.m) et **SE sémantique autre** qui regroupe les SE qui ne sont ni à visée ontologique, ni métalinguistiques (fig. 3.o).

Une typologie des relations est associée aux types sémantiques "à visée ontologique" et "métalinguistique". Les relations *is-a* (fig. 3.a, fig 3.b, fig. 3.c), *part-of* (fig. 3.d, fig. 3.g), *instance-of* (fig. 3.n), *ontologique autre* (relation ontologique transverse ou d'actance) (fig. 3.i) sont associées aux SE à visée ontologique.

Les relations d'*hyperonymie*, de *méronymie*, d'*homonymie* (fig. 3.m), de *synonymie*, de *multilinguisme* (fig. 3.l), *lexicale autre* (relation lexicale moins fréquente décrivant la

langue, telle que la paronymie qui associe deux mots à la graphie/prononciation proches mais aux sens différents) sont associées aux SE métalinguistiques.

De façon orthogonale, les connaissances portées par la SE peuvent être contextualisées dans l'espace (fig. 3.j, fig. 3.n), dans le temps, ou dans tout autre dimension (fig. 3.m), à l'aide de circonstants. L'annotation de ces derniers permet d'envisager l'identification de relations autres que binaires. Nous distinguons les **SE contextuelles** des **SE non contextuelles**.

(a) Deux phénomènes sont responsables de l'augmentation substantielle du rayon de l'étoile (qui peut atteindre un rayon 1 000 fois supérieur à celui du Soleil). Premièrement, la fusion en couche de l'hydrogène. Et deuxièmement, la contraction du cœur d'hélium, libérant une importante quantité d'énergie gravitationnelle.

(b) Le dromadaire a été répertorié dans 35 pays, tels que l'Inde, la Turquie, le Kenya, le Pakistan, la corne de l'Afrique et bien d'autres encore.

(c) Les Grecs fabriquent généralement des meubles en bois (type érable, chêne, if, saule), mais aussi en pierre et en métal (bronze, fer, or, argent).

(d) Une chaussure se compose principalement :
- du semelage, partie qui protège la plante des pieds, plus ou moins relevée à l'arrière par le talon
- de la tige, partie supérieure qui enveloppe le pied

(e) Le bénéfice imposable est la différence entre les recettes et les charges de l'entreprise durant l'exercice comptable.
- Sont pris en compte pour les produits (recettes) :
  - les produits d'exploitation autrement dit le chiffre d'affaires de l'entreprise ;
  - les produits accessoires, c'est-à-dire les recettes.
- Sont pris en compte pour les charges (. . .) retenues pour leur coût hors taxe :
  - les frais généraux : salaire, loyer commercial, frais de bureau, etc. ;
  - les charges financières (agios, intérêts d'emprunt)

(f) Est considéré comme "lecture savante", du point de vue fonctionnel, une pratique de lecture répondant aux critères suivants :
- c'est une lecture "qualifiée",
- qui se développe sur le temps long de la recherche scientifique,
- dans un parcours forcément individualisé,
- où l'écriture se combine à la lecture, souvent dans une perspective de publications.

(g) Chaque nucléotide est constitué de trois éléments liés entre eux :
- un groupe phosphate lié à :
- un sucre, le désoxyribose, lui-même lié à :
- une base azotée.

(h) Préparation de la recette :

Lavez les asperges, épluchez-les de la pointe vers la base. Faites-les cuire dans une casserole d'eau bouillante avec les tablettes de bouillon pendant 25 à 30 minutes.

Égouttez-les et déposez-les précautionneusement sur du papier absorbant. Laissez-les refroidir.

Coupez-les en deux en réservant les pointes d'une longueur de 10 à 12 cm d'une part, les queues d'autre part.

(i) Selon ce décret, la BnF a pour mission :
- de collecter, cataloguer, conserver et enrichir dans tous les champs de la connaissance, le patrimoine national dont elle a la garde, en particulier le patrimoine de langue française ou relatif à la civilisation française.
- d'assurer l'accès du plus grand nombre aux collections, sous réserve des secrets protégés par la loi, dans des conditions conformes à la législation sur la propriété intellectuelle et compatibles avec la conservation de ces collections.

(j) Les Berbères ont mené une vive résistance parfois qualifiée de "farouche".
- Algérie : De nombreux soulèvements ont été menés pour contrer la colonisation française, l'émir Abd el-Kader qui faisait remonter ses origines à la tribu berbère des Banou Ifren (Zénètes) a lutté après avoir déclaré la guerre aux Français, il fut capturé puis fait prisonnier. En juillet 1857, (...)
- Maroc : Le mouvement de résistance s'est illustré lors de la guerre du Rif menée par Abdelkrim al-Khattabi, qui est une guerre coloniale qui opposa les tribus berbères du rif aux armées françaises et espagnoles, de 1921 à 1926. (...)
- Libye : La lutte contre la colonisation italienne est d'abord menée par Omar Al Mokhtar surnommé "Cheikh des militants " qui est un chef musulman libyen d'origine berbère qui organisa la lutte armée contre la colonisation italienne au début du XXe siècle. D'autres leaders nationalistes (...)

(k) Du point de vue de la tradition textuelle juive, la division en chapitres est non seulement une innovation étrangère sans aucun fondement dans la messora, mais elle est également fort critiquable car :
- la division en chapitres reflète souvent l'exégèse chrétienne de la Bible ;
- quand bien même ce ne serait pas le cas, elle est artificielle, divisant le Texte en des endroits jugés inappropriés pour des raisons littéraires ou autres.

(l) Munich [mynik] (München en allemand, Minga en bavarois) est, avec 1 443 122 habitants1, la troisième ville d'Allemagne par la population après Berlin et Hambourg.

(m) Une arête est un nom commun féminin qui peut désigner :
- l'arête, 'barbe de l'épi de graminées' (notion de botanique) ;
- l'arête, 'partie du squelette d'un poisson' (notion d'ichtyologie) ;
- l'arête, 'ligne d'intersection de deux plans' (notion de géométrie dans l'espace, d'architecture, etc.).

(n) Manoirs célèbres
- Le manoir d'Ango à Varengeville-sur-mer, près de Dieppe.
- Le manoir de Brion au Mont-Saint-Michel
- Le manoir d'Eyrignac à Salignac-Eyvigues en Périgord

(o) S sait que p si et seulement si
1. p est vrai ;
2. S croit que p ; et
3. la croyance de S dans p est justifiée.

Figure 3: Exemples de SE issus de pages Wikipedia ou du corpus de Virbel (1999)

4 Processus d'annotation

La typologie décrite ouvre la voie à une caractérisation plus fine des SE. Corollaire de cette possibilité, elle offre une latitude plus large pour la discrimination des classes lors d'un apprentissage supervisé pour l'identification des relations que portent les SE (Fauconnier et al., 2013).

Afin d'éprouver cette typologie de manière empirique, nous avons débuté une campagne d'annotation avec trois annotateurs. La tâche d'annotation elle-même se déroule en trois phases principales qui consistent à :

(1) délimiter les différents composants de la SE (amorce, items, clôture) lorsqu'elle bénéficie de mise en forme.

(2) annoter la SE selon les critères rhétoriques, intentionnels et sémantiques définis ci-dessus. Chaque SE se voit affecter un type rhétorique, un ou plusieurs types intentionnels, un type sémantique. Lorsque la SE est paradigmatique, à visée ontologique ou métalinguistique, un type de relation est associé au type sémantique (associée ou non à un contexte).

(3) délimiter, lorsque la SE est paradigmatique et à visée ontologique ou métalinguistique, les unités textuelles qui dénotent le concept présent dans l'amorce, le concept présent dans chacun des items, le circonstant (lorsqu'il existe) et la relation entre l'amorce et chacun des items.

Pour être menée à bien, cette tâche d'annotation nécessitait un outil adapté à la caractérisation multi-dimensionnelle des SE, cas moins courant en TAL où l'on privilégie habituellement des annotations simple label. De plus, il était aussi indispensable que cet outil supporte le caractère imbriqué et potentiellement récursif des SE. Par exemple, une SE peut contenir d'autres SE et elle-même être imbriquée au sein d'une structure discursive plus large (e.g : citation) ou être étalée sur plusieurs d'entre elles (e.g : un titre et plusieurs paragraphes). Enfin, cet outil devait être modulable pour être facilement adapté à d'autres types d'objets avec mise en forme (e.g : énoncés définitoires, démonstrations mathématiques, etc.) et plusieurs types de format d'entrée (e.g : HTML, PDF, etc.).

Les outils d'annotation tels que MMAX2 (Müller and Strube, 2006), MAE (Stubbs, 2011) ou encore Glozz (Widlöcher and Mathet, 2009) ne répondent pas ou partiellement à ces exigences. MMAX2 et MAE prennent du texte brut en entrée et ne gardent pas la mise en forme originelle des textes. Glozz, initialement conçu pour l'annotation de relations discursives, supporte la mise en forme du texte mais n'est, en l'état, pas adapté pour une annotation rapide et ergonomique d'objets multi-labels. En outre, la possibilité de faire évoluer le code source de Glozz n'est pas assurée (licence restrictive).

Pour toutes ces raisons, nous avons développé LARAt (Logiciel d'Acquisition de Relations par l'Annotation de textes[5]), prononcé /laʁa/. Cet outil Java se veut portable, et open-source. Dans son état actuel, LARAt prend en entrée des fichiers HTML ou XML respectant la norme TEI[6], les affiche en respectant leur mise en forme et permet aux annotateurs d'annoter des objets textuels imbriqués ou éclatés sur plusieurs niveaux textuels (e.g : titres et sous-titres).

Dans la tâche d'annotation des SE, deux types d'annotation sont produits (type 1 et type 2). Les annotations de type 1 concernent exclusivement le repérage en document des SE. Une fois délimitée, les SE sont caractérisées avec des annotations de type 2 qui reprennent les éléments décrits dans la typologie présentée. Ainsi, à chaque annotation de type 1 est associée une ou plusieurs annotations de type 2. Cette manière modulaire de gérer l'annotation facilite les post-traitements et l'emploi spécialisé de ces dernières (e.g : étude d'un phénomène particulier, recherche d'un cas précis pour exemplifier un emploi, etc.).

À terme, cet outil sera amené à supporter le PDF ainsi que le post-traitement des annotations (alignement, Kappa de Cohen et Fleiss pour l'accord inter-annotateurs).

À noter qu'un guide d'annotation accompagne cette campagne d'annotation. Sa rédaction se déroule de manière itérative en prenant en compte les retours des annotateurs et les cas ambigus qui posent question. Au terme de la campagne, le corpus annoté, le guide ainsi que LARAt seront distribués sous licence libre.

[5](en) *Layout Annotation for Relations Acquisition tool*

[6]Text Encoding Initiative

5 Conclusion et perspectives

L'analyse que nous avons menée sur les SE a permis de définir une typologie multi-dimensionnelle, permettant de tenir compte de propriétés de nature différente et parfois orthogonales. Le but théorique de ce travail a été d'élucider le phénomène complexe des SE quant à sa forme, sa structure ou sa fonction. D'un point de vue pratique, ce travail nous permet d'une part d'améliorer le repérage des SE dans les textes et, d'autre part d'identifier la ou les relations sémantiques qui relient les concepts contenus dans la SE. À cet égard, nous avons développé l'outil d'annotation LARAt qui permet de catégoriser les SE extraites de textes suivant les différents axes de notre typologie. Une première campagne d'annotation à l'aide de cet outil est en cours. La principale perspective de poursuite de ce travail est son extension à d'autres objets textuels ayant un impact sur la sémantique des textes tels que la titraille et les énoncés définitoires.

Références

N. Aussenac-Gilles and M.-P. Jacques. 2008. Designing and evaluating patterns for relation acquisition from texts with Caméléon. *Terminology*, 14:45–73.

P. Buitelaar, P. Cimiano, and B. Magnini. 2005. Learning taxonomic relations from heterogeneous sources of evidence. In P Buitelaar, P Cimiano, and B Magnini, editors, *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123, pages 59–73. IOS Press, Amsterdam.

J. Fauconnier, M. Kamel, B. Rothenburger, and N. Aussenac-Gilles. 2013. Apprentissage supervisé pour l'identification de relations sémantiques au sein de structures énumératives parallèles. In *Actes de la 20e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, pages 132–145.

Z. Harris. 1976. A theory of language structure. *American Philosophical Quarterly*, 13(4):237–255.

M. A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, volume 2, pages 539–545. Association for Computational Linguistics.

L.-M. Ho-Dac, M.-P. Péry-Woodley, and L. Tanguy. 2010. Anatomie des structures énumératives. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*.

E. H. Hovy and Y. Arens. 1991. Automatic Generation of Formatted Text. In *Proceedings of the 9th AAAI Conference (AAAI 1991)*, Anaheim, CA.

M. Kamel and B. Rothenburger. 2011. Elicitation de Structures Hiérarchiques à partir de Structures Enumératives pour la Construction d'Ontologie. In *Journées Francophones d'Ingénierie des Connaissances (IC 2011)*, pages 505–522, Annecy.

C. Luc. 2001. Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. In *Actes de la 8e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, pages 263–272.

W.C. Mann and S.A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

E. Montiel-Ponsoda and G. A. de Cea. 2011. Using natural language patterns for the development of ontologies. In V. Bhatia, P. Sánchez Hernández, and P. Pérez Paredes, editors, *Researching specialized languages*, volume 47, pages 211–230. John Benjamins.

C. Müller and M. Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In S. Braun, K. Kohn, and J. Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

E. Pascual. 1991. *Représentation de l'architecture textuelle et génération de texte*. Ph.D. thesis, Université Paul Sabatier. Toulouse, France.

J. Poelmans, P.l Elzinga, S. Viaene, and G. Dedene. 2010. Formal concept analysis in knowledge discovery: a survey. In M. Croitoru, S. Ferré, and D. Lukose, editors, *Conceptual Structures: From Information to Intelligence*, volume 18, pages 139–153. Springer.

S. Porhiel. 2007. Les structures énumératives à deux temps. *Revue romane*, 42(1):103–135.

A. Stubbs. 2011. MAE and MAI: Lightweight Annotation and Adjudication Tools. In *2011 Proceedings of the Linguistic Annotation Workshop V, Association of Computational Linguistics*, Portland.

M. Vergez-Couret, L. Prévot, and M. Bras. 2008. Interleaved discourse, the case of two-step enumerative structures. In *Proceedings of Contraints In Discourse III*, pages 85–94, Potsdam.

J. Virbel. 1989. The contribution of linguistic knowledge to the interpretation of text structures. pages 161–180.

J. Virbel. 1999. Structures textuelles, planches fascicule 1 : Enumérations, Version 1,. Technical report, IRIT.

A. Widlöcher and Y. Mathet. 2009. La plateforme Glozz : environnement d'annotation et d'exploration de corpus. In *Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*.