# Analyse comparative de méthodologies et d'outils de construction automatique d'ontologies à partir de ressources textuelles

Toader Gherasim, Mounira Harzallah, Giuseppe Berio, Pascale Kuntz

# Analyse comparative de méthodologies et d'outils de construction automatique d'ontologies à partir de ressources textuelles

Toader Gherasim*, Mounira Harzallah*,
Giuseppe Berio**, Pascale Kuntz *

*LINA, UMR 6241 CNRS
toader.gherasim, mounira.harzallah @univ-nantes.fr, pascale.kuntz@polytech.univ-nantes.fr,
http://www.lina.univ-nantes.fr
**LABSTICC, UMR 3192 CNRS
giuseppe.berio@univ-ubs.fr
http://www.lab-sticc.fr

**Résumé.** Plusieurs méthodologies et outils de construction automatique des ontologies à partir de ressources textuelles ont été proposés ces dernières années. Dans cet article nous analysons quatre approches en les comparant à une approche de référence – Methontology. Dans leur sélection nous avons privilégié celles qui couvrent l'ensemble des étapes du processus de construction d'ontologies. Puis nous analysons et comparons la portée, les limites et les performances des implémentations logicielles associées aux approches analysées. Ces outils ont été testés sur un corpus de ressources textuelles, et nous avons comparé leurs résultats à ceux obtenus manuellement.

## 1 Introduction

Depuis les travaux fondateurs de Gruber (1993), les ontologies jouent un rôle majeur en ingénierie des connaissances, et leur essor, associé à celui du web sémantique, ne cesse de croître. Elles sont maintenant au coeur des systèmes de recherche d'information ou d'aide à la décision de multiples domaines (e.g. médical Osborne et al. (2009), juridique Bourigault et Lame (2002), etc.). Initialement de tailles restreintes et construites entièrement "à la main", elles peuvent contenir aujourd'hui plusieurs milliers de concepts (e.g. Aime et al. (2009)) et des relations variées. Et, si dans la lignée des systèmes experts, leur construction a reposé, et repose encore, sur l'expertise humaine, leur popularisation et le changement d'échelle nécessitent de plus en plus le recours massif à une forme d'automatisation qui limite la contribution des experts. L'accessibilité croissante de nombreuses ressources textuelles qui renferment des connaissances d'une grande richesse sur les concepts et les processus mis en oeuvre dans les applications rend cette aspiration de plus en plus crédible.

Des propositions à la fois méthodologiques et logicielles se sont multipliées ces dernières années pour construire automatiquement des ontologies à partir de corpus textuels (Velardi et al. (2007), Maynard et al. (2009b)). Développées souvent dans des contextes applicatifs, ce

qui a permis leur validation, ces approches proposent différentes techniques et ingénieries pour l'automatisation des tâches qui composent le processus de construction d'ontologies.

L'objectif de cet article est de proposer une analyse comparative de certaines de ces approches et des outils associés selon deux directions complémentaires : technique et expérimentale. Afin de donner un cadre cohérent à cette analyse, nous avons utilisé Methontology (Fernandez et al., 1997), qui est une méthodologie largement diffusée et reconnue au sein de la communauté de l'ingénierie des ontologies, comme référentiel des tâches nécessaires à la construction d'une ontologie.

Dans cet article, nous avons considéré quatre approches : OntoLearn (Navigli et al. (2003), Velardi et al. (2007)), Text2Onto (Cimiano et Volker (2005), Volker et Blomqvist (2008)), Alvis (Nedellec, 2006) et SPRAT (Maynard et al., 2009a,b), et 6 outils qui leur sont associés : *Text2Onto*, *TermExtractor*, *GlossExtractor*, *SSI*, *TermRaider* et *SPRAT*. Dans notre choix nous avons privilégié les approches "complètes", qui ne sont pas seulement des aides à la construction ou des éditeurs, mais qui couvrent toute la démarche, de l'analyse textuelle à la proposition d'une première ébauche d'ontologie (concepts, relations et instances).

Dans la première partie de l'article, nous rappelons la composition du référentiel de tâches de l'activité de conceptualisation de Methontology et nous positionnons par rapport à ce référentiel les quatre approches considérées et les outils qui leur sont associés. Dans la deuxième partie, nous proposons une analyse des six outils logiciels cités précédemment. Nous les avons comparés d'abord selon des critères techniques (types de données d'entrée, contraintes d'utilisation, etc.), puis nous avons comparé, sur un corpus textuel, leurs sorties avec des résultats obtenus par une construction manuelle suivant les tâches de Methontology. Les résultats des outils s'avèrent prometteurs mais ils doivent être souvent raffinés ou complétés. Dans la discussion qui conclut l'article nous comparons les quatre approches en fonction de l'utilisation (ou de la non utilisation) des ressources externes.

# 2 Méthodologies de construction

## 2.1 "Methontology"

L'approche Methontology est décrite par ses auteurs comme générale et indépendante du champ applicatif (Fernandez et al. (1997), Corcho et al. (2005)). La conceptualisation des connaissances, qui sert de support à l'ontologie, est structurée autour de sept tâches. La figure 1 indique un certain ordre dans l'exécution des tâches, ordre qui doit être suivi pour assurer la consistance et la complétude des connaissances représentées. Cependant, le processus de conceptualisation dans son ensemble n'est pas séquentiel, un retour vers les tâches réalisées précédemment est possible à tout moment.

La **tâche 1** est la construction d'un glossaire de termes qui doit contenir toutes les connaissances du domaine qui sont utiles et potentiellement utilisables pour la construction de l'ontologie de ce domaine. Ce glossaire comprend des concepts, des instances, des verbes et des attributs. La **tâche 2** porte sur la construction des taxonomies de concepts. Les termes synonymes, qui représentent un concept, sont tout d'abord regroupés sous un seul intitulé qui sert d'étiquette pour le concept puis, les concepts "étroitement liés" sont regroupés par catégories. Les concepts sont considérés comme "étroitement liés" entre eux s'ils peuvent être tous structurés dans une taxonomie sans faire appel à des nouveaux concepts, plus abstraits ou généraux,

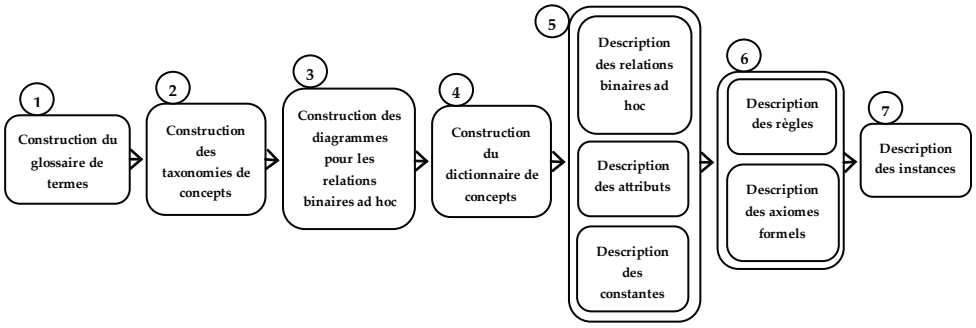


FIG. 1 – *Les tâches composant l'activité de Conceptualisation de Methontology (schéma emprunté de (Corcho et al., 2005)).*

qui ne sont pas directement obtenus à partir de la conceptualisation des termes présents dans le glossaire. La **tâche 3** définit la construction des diagrammes pour les relations binaires ad hoc. Ces relations sont identifiées à partir de syntagmes verbaux. Le nom de la relation est identifié à l'aide du regroupement des verbes synonymes sous un seul intitulé. La relation porte sur les concepts identifiés dans l'étape précédente correspondant au sujet et au complément d'objet présents dans le syntagme verbal. La **tâche 4** est la construction d'un dictionnaire de concepts dans lequel sont regroupées toutes les informations relatives aux concepts (leur sémantique, leurs attributs, leurs instances, etc.). La **tâche 5** permet une description détaillée des relations, des attributs des concepts et des constants précédemment identifiés. La **tâche 6** concerne la description des règles et des axiomes formels portant sur les divers éléments de l'ontologie déjà connus - les concepts, les instances, les constants, etc. La **tâche 7** concerne la description détaillée des instances. Les tâches 5, 6 et 7 peuvent être considérées comme des tâches de raffinement de l'ontologie suite à l'identification des concepts et des relations.

## 2.2 Comparaisons

Dans ce paragraphe nous utilisons Methontology comme un cadre de référence pour situer et comparer sur des bases communes les diverses tâches et sous-tâches des quatre approches sélectionnées (OntoLearn, Alvis, Text2Onto et SPRAT) et pour identifier les techniques utilisées pour effectuer chacune de ces tâches.

Les quatre approches se focalisent sur les aspects liés à la construction de la structure de l'ontologie, aspects présents surtout dans les quatre premières tâches de Methontology (construction du glossaire de termes, des taxonomies de concepts, des diagrammes des relations binaires et du dictionnaire de concepts), et considèrent peu les aspects de raffinement présents dans les tâches 5, 6 et 7 de Methontology. Cependant, l'équivalence entre les tâches de Methontology et les tâches des quatre approches n'est pas complète et on peut noter au moins deux différences importantes.

La première porte sur les tâches elles-mêmes. Un certain nombre de sous-tâches qui composent les tâches de Methontology n'ont pas toujours une sous-tâche équivalente dans le cadre des quatre approches ; c'est le cas, par exemple, de la sous-tâche de définition des termes, pré-

sente dans OntoLearn mais absente des trois autres approches. De plus, même si une sous-tâche équivalente existe, elle peut avoir un objectif un peu différent ; la sous-tâche d'identification de termes dans Methontology concerne les concepts, les instances et les attributs, dans OntoLearn la sous-tâche équivalente concerne seulement les concepts, et dans les trois autres approches les sous-tâches concernent les concepts et les instances.

La deuxième différence porte sur la séparation entre les tâches et leur ordonnancement. L'identification de concepts et l'identification des relations sont deux tâches distinctes dans Methontology et Text2Onto et les concepts sont identifiés avant les relations. En revanche, dans l'approche SPRAT, les concepts et une partie des relations qui leur sont associées sont découverts conjointement, dans le cadre d'un processus itératif.

Le tableau 1 synthétise les principales correspondances entre les tâches de l'activité de conceptualisation de Methontology et celles des approches considérées. Il présente également les techniques utilisées par chaque approche pour la réalisation des tâches. Le niveau de détail dans la description des sous-tâches diffère entre approches (e.g. entre OntoLearn et Alvis), et cela parce que la complexité et le niveau de structuration diffèrent d'une approche à l'autre.

# 3 Analyse comparative des outils pour la construction semi-automatisée d'ontologies à partir de ressources textuelles

D'une façon générale, nous pouvons regrouper les outils utilisés en support des approches décrites dans la section 2 en deux grandes classes : une classe comprenant des outils génériques, qui ont été développés dans un autre but que la construction ou l'aide à la construction des ontologies et qui ont trouvé une utilisation particulière pour cette problématique, et une classe comprenant des outils consacrés spécifiquement à la construction d'ontologies.

La première classe comprend notamment des outils d'apprentissage inductif de règles (*C4.5* (Navigli et al., 2003) pour OntoLearn et *Propal* (Nedellec, 2006) pour Alvis), des analyseurs grammaticaux (*LINKPARSER* et *BioLG* (Nedellec, 2006) pour Alvis), des extracteurs de termes (*YATEA* (Nedellec, 2006) pour Alvis) ou des outils de traitement de patrons textuels (*JAPE* (Maynard et al., 2009a,b) pour SPRAT). La deuxième classe comprend des outils qui sont notamment associés aux méthodologies présentées ci-dessus : *Text2Onto* (Cimiano et Volker, 2005) pour l'approche Text2Onto, *TermExtractor*, *GlossExtractor* (Velardi et al., 2008) et *SSI* (Navigli et Velardi, 2005) pour l'approche OntoLearn, *NEBOnE*, *TermRaider* et *SPRAT* (Maynard et al., 2009a) pour l'approche SPRAT et *ASIUM* (Nedellec, 2006) pour l'approche Alvis. A notre connaissance, *NEBOnE* et *ASIUM* ne sont pas disponibles pour être testés.

Le tableau 2 présente les correspondances qui existent entre ces outils et les diverses tâches et sous-tâches des quatre approches. Certains outils couvrent une seule sous-tâche de l'approche associée (ex. *GlosssExtractor*) alors que d'autres peuvent en couvrir plusieurs ou même la totalité (e.g. *SPRAT*). *JAPE* et *NEBOnE* servent de support à *TermRaider* et à *SPRAT* mais ne peuvent pas réaliser seuls les sous-tâches en question. Le tableau 2 permet ainsi d'identifier les outils qui peuvent être comparés entre eux : par exemple, il n'est pas envisageable de comparer *LINKPARSER* avec *TermRaider* mais il est possible de comparer *TermRaider* avec *TermExtractor* et *Text2Onto*.

| Les tâches de Methontology | Les tâches d'OntoLearn | Les tâches d'Alvis | Les tâches de Text2Onto | Les tâches de SPRAT |
|---|---|---|---|---|
| **Construction du glossaire de termes**<br>*Identification et définition des termes correspondant à des concepts, à des verbes, à des instances et à des attributs* | Extraction des termes – *technique mixte (linguistique et statistique)*<br>Filtrage des termes – *technique mixte (utilisant indices statistiques et linguistiques)*<br>Validation des termes – *manuellement*<br>Identification des définitions pour les termes *(obtenant ainsi un glossaire) – à l'aide de recherches sur Internet*<br>*Les termes concernent seulement des concepts* | Extraction des termes – *technique linguistique*<br>Validation des termes – *manuellement*<br><br>*Les termes concernent seulement des concepts et des instances* | Extraction de termes – *technique mixte*<br><br>*Les termes concernent seulement des concepts et des instances* | Extraction de termes – *technique linguistique*<br>*Les termes concernent seulement des concepts et des instances*<br>**<u>Exécution itérative</u>**<br>Choix d'un nouveau terme qui est relié par un ou des patrons (correspondant à *'is a'*) à la racine de l'ontologie ou à des concepts déjà connus |
| **Construction des taxonomies de concepts**<br>*Regroupement des termes synonymes sous le même intitulé*<br>*Regroupement des concepts liés par des relations 'is-a' dans une même taxonomie* | Construction d'arbres lexicaux *(en utilisant la structure des termes composés) – technique structurelle*<br>Interprétation sémantique des termes *(pour identifier des nouvelles relations taxonomiques entre les termes composant les arbres) – en utilisant un algorithme (SSI) spécifique à OntoLearn, qui utilise WordNet*<br>Réorganisation des arbres lexicaux en arbres conceptuels – *en utilisant les relations taxonomiques identifiées précédemment – manuellement* | Construction d'une taxonomie – *Identification des attributs contextuels de chaque terme ; construction de la taxonomie en utilisant les attributs pour la classification non supervisée des termes et des classes de termes (concepts) – technique distributionnelle* | Identification des relations *'is a'* entre concepts *(à l'aide de WordNet, des patrons et des heuristiques (portant sur les termes composés) – technique mixte (structurelle et contextuelle)* | *(si le terme choisi est un concept)*<br>Insertion du concept dans la structure taxonomique de l'ontologie – *à l'aide des règles de création, d'insertion et de gestion des conflits, des règles spécifiques à SPRAT* |
| **Construction des diagrammes des relations binaires ad hoc**<br>*Regroupement des verbes synonymes sous le même intitulé*<br>*Définition des relations à partir de verbes en rajoutant à chaque verbe des couples de concepts : sujet et complément d'objet* | Sélection d'une liste de relations considérées importantes pour le domaine concerné – *manuellement*<br>Sélection de quelques exemples *(impliquant les relations choisies et quelques uns des concepts identifiés précédemment)* ; description des éléments composant les exemples à l'aide des vecteurs de caractéristiques – *manuellement*<br>Apprentissage des règles pour l'identification des relations – *programme d'apprentissage inductif de règles qui utilise la description des exemples*<br>Identification des concepts non impliqués dans les exemples mais qui sont reliés par les relations choisies – *automatiquement, à l'aide des règles* | Analyse syntaxique détaillée du texte – *utilisant ASA, formalisme spécifique à Alvis*<br>Sélection des quelques exemples (des phrases qui contient des relations) – *manuellement*<br>Apprentissage de règles pour l'identification des relations – *programme d'apprentissage inductif de règles qui utilise ASA*<br>Identification des concepts reliés par les relations présentes dans les exemples – *automatiquement, à l'aide des règles* | Identification des relations de relativité *('subtopic of')* entre concepts – *(à l'aide de l'analyse statistique des co-occurrences des concepts) – technique distributionnelle*<br>Identification des relations générales (définies par des verbes) – *(à l'aide de l'analyse des syntagmes verbaux et de la fréquence d'apparition des concepts dans les syntagmes centrées autour d'un même verbe) – technique distributionnelle* | Identification et insertion dans l'ontologie des relations entre le nouvel concept et des autres concepts déjà connus *(les relations correspondent à des patrons prédéfinis) – technique contextuelle* |
| **Construction du dictionnaire de concepts** *(contenant, pour chaque concept, son sens, ses relations, ses attributs et ses instances)* | *Le sens des concepts a été déjà identifié pendant l'interprétation sémantique de termes* | — | Identification des relations du type *'instance of' – technique distributionnelle* | *(si le terme choisi est une instance)*<br>Insertion de l'instance dans l'ontologie – *à l'aide des règles spécifiques à SPRAT* |

TAB. 1 – *Correspondances entre les quatre premières tâches de l'activité de conceptualisation de Methontology et celles des approches analysées.*

## 3.1 Comparaison technique

Dans les paragraphes qui suivent nous avons comparé les six outils disponibles selon cinq critères. Ainsi, en ce qui concerne **le type et les formats des entrées** les quatre outils qui permettent l'analyse du texte (*TermExtractor*, *Text2Onto*, *TermRaider* et *SPRAT*) prennent en entrée des fichiers sans structure particulière (du type *'.txt'*). Seul *TermExtractor* supporte des autres formats (PDF, DOC, HTML, etc. ou archives contenant ce type de fichiers). *TermExtractor* est le seul capable d'intégrer dans son analyse des aspects concernant les styles textuels (gras, italique, etc.). Et *GlossExtractor* et *SSI* prennent en entrée des listes de termes.

**En sortie** *TermExtractor* et *TermRaider* offrent une liste de termes et *GlossExtractor* une liste de définitions. *SSI* construit un réseau sémantique, *SPRAT* une ontologie et *Text2Onto* permet d'obtenir des listes de concepts, d'instances et de relations ou directement une ontologie.

La principale **contrainte technique d'utilisation** concerne la possibilité d'installer ces outils en régie propre. En fait, la majorité des outils (*TermExtractor**, *GlossExtractor**, *SSI**, *TermRaider*** et *SPRAT***) sont disponibles sur les serveurs de leurs créateurs, soit via une interface web*, soit sur la forme d'un service web**, et, de ce fait, le volume de données traitables est limité. Cette limite concerne notamment *SSI* et *GlossExtractor* qui acceptent des listes de maximum dix termes (termes simples – un seul mot – dans le cas de *SSI*). *Text2Onto* est le seul outil qui peut être installé en régie propre.

La majorité des outils font appel à des **connaissances exogènes** comme WordNet (*Text2Onto*, *SSI*), des dictionnaires (*GlossExtractor*, *Text2Onto*, *SPRAT*) ou des recherches sur internet (*GlossExtractor*, *Text2Onto*). *TermExtractor* et *TermRaider* sont les seuls qui peuvent fonctionner sans utiliser des ressources externes.

Seuls deux outils - *TermExtractor* et *Text2Onto* - permettent des **configurations**. *TermExtractor* propose des configurations concernant notamment la prise en charge des styles textuels, les valeurs limites pour les indices utilisés dans le processus de filtrage des termes et le nombre maximal de termes qui peuvent faire partie d'un terme composé. *Text2Onto* propose plusieurs algorithmes pour l'identification des divers éléments de l'ontologie ; ce qui permet de combiner leurs résultats.

Trois outils - *TermExtractor*, *Text2Onto* et *SPRAT* - peuvent **faire évoluer leurs résultats**. *TermExtractor* peut enrichir une liste de termes identifiés précédemment à partir des nouvelles ressources textuelles. *Text2Onto* est capable de gérer l'évolution des ressources d'entrée en faisant évoluer d'une manière dynamique les résultats déjà obtenus. *SPRAT* peut prendre en entrée une ontologie et l'enrichir avec des connaissances extraites à partir de ressources textuelles, ou bien il peut en construire une.

## 3.2 Comparaison expérimentale

Pour mieux cerner la pertinence de ces outils, nous avons réalisé une étude où nous les avons testés sur un volume réduit de ressources textuelles et nous avons comparé leurs résultats à ceux que nous avons obtenus manuellement. Les ressources textuelles choisies portent sur le domaine de la construction d'ontologies. Elles sont constituées de courts extraits d'articles scientifiques, contenant au total 508 mots. Les différents résultats obtenus et les comparaisons que nous avons faites sont disponibles à la demande. L'étude réalisée nous a permis de prendre en main ces différents outils et de réaliser une première comparaison de la pertinence de leurs résultats.

### 3.2.1 Analyse des résultats des outils qui identifient des termes et des définitions

Nous avons testé sur les mêmes ressources textuelles les quatre outils qui identifient des termes - *TermExtractor*, *TermRaider*, *SPRAT* et *Text2Onto*. Notons que nous avons obtenu des résultats différents pour *TermRaider* et *TermExtractor* selon que les ressources textuelles étaient regroupées dans un seul fichier ou dans plusieurs (un fichier pour chaque extrait d'article). En revanche *Text2Onto* et *SPRAT* fournissent des résultats identiques.

Le tableau 3 présente : le nombre de termes identifiés par chaque outil et manuellement (suivant les tâches 1 et 2 de Methontology), la structure des termes identifiés (le nombre de mots qui composent le terme) et le nombre (et le pourcentage) de termes bien identifiés. Pour les rubriques *'Manuel'* et *'Text2Onto'* nous avons précisé le nombre de concepts et d'instances parmi les termes extraits (entre parenthèses). Pour *TermRaider* et pour *TermExtractor* nous avons distingué les deux cas de test (un seul ou plusieurs fichiers d'entrée). On remarque que *Text2Onto* a obtenu le meilleur rappel, identifiant 73% des termes identifiés manuellement, mais une faible précision de seulement 28%. *TermExtractor* sur plusieurs fichiers a une très bonne précision (75%) mais un rappel de 10%. *TermRaider* sur plusieurs fichiers et *TermExtractor* sur un seul fichier se trouvent entre ces deux cas extrêmes.

Comme *Text2Onto* identifie des concepts et des instances, et non des termes, il est possible que certains termes identifiés en tant que concepts par la construction manuelle soient identifiés en tant qu'instances par *Text2Onto* et vice-versa. Nous avons ainsi complété l'analyse présentée dans le tableau 3 en dissociant les cas : termes identifiés en tant que concepts (resp. instances) par l'analyse manuelle et en tant que concepts (resp. instances) par *Text2Onto*. La majorité (25 sur les 27 en commun) des résultats communs entre *Text2Onto* et la construction manuelle sont également considérés comme des concepts (resp. instances). Seuls un concept et une instance issus de la construction manuelle ont été identifiés en tant qu'instance (resp. concept) par *Text2Onto*.

*SPRAT* n'identifie pas simplement une liste de termes. Dès qu'il identifie un terme il construit le concept et les relations correspondantes et les intègre dans l'ontologie en cours de construction. Testé sur nos ressources textuelles *SPRAT* a fournit une ontologie vide, ce qui explique les valeurs de zéro dans la colonne correspondant à *SPRAT* dans le tableau 3. Ce résultat est dû, sans aucune doute, à la taille trop réduite de nos ressources. Nous avons testé *SPRAT* sur des ressources textuelles du même type mais plus volumineuses (autour de 20000 mots), et *SPRAT* a construit une ontologie très simple contenant seulement 14 concepts.

Pour vérifier la pertinence des résultats obtenus par l'outil *GlossExtractor* nous l'avons testé sur les 18 termes identifiés par *TermExtractor*. L'outil a identifié 83 définitions qui concernent seulement 9 de ces termes. Pour les 9 autres termes l'outil n'a identifié aucune définition. Parmi les 83 définitions il y a un certain nombre de doublures (la même définition associée plusieurs fois au même terme), quelques définitions circulaires (qui définissent le terme par lui même) ou encore des phrases qui ne constituent pas une définition. De plus, comme GlossExtractor ne permet pas de spécifier un domaine d'intérêt, les définitions obtenues couvrent tous les sens des termes. Un filtrage de ces définitions s'est donc imposé et à la fin nous avons retenu seulement 30 définitions qui concernent 8 termes. Finalement *GlossExtractor* a identifié des définitions pour 50% des termes et seulement 35% des définitions ont été validées. Les définitions validées peuvent être le point de départ pour l'identification automatique des relations taxonomiques.

| **Les tâches de Methontology** | **Les tâches d'OntoLearn** | | **Les tâches d'Alvis** | | **Les tâches de Text2Onto** | | **Les tâches de SPRAT** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Sous-tâche** | **Outil** | **Sous-tâche** | **Outil** | **Sous-tâche** | **Outil** | **Sous-tâche** | **Outil** | | |
| **Construction du glossaire de termes** | Extraction des termes | **Term Extractor** | Extraction des termes | *YATEA* | Extraction de termes (concepts) | **Text2Onto** *m. Concept* | Extraction des termes | *JAPE, NEBONE* | **Term Raider** | **SPRAT** |
| | Filtrage des termes | | | | | | | | | |
| | Validation des termes | — | Validation des termes | — | Extraction de termes (instances) | **Text2Onto** *module Instance* | Choix d'un terme | | — | |
| | Identification des définitions | **Gloss Extractor** | | | | | | | | |
| **Construction des taxonomies de concepts** | Construction d'arbres lexicaux | — | Construction d'une taxonomie | *BioLG, ASIUM implémenté pour Alvis* | Identification des relations *'is a'* | **Text2Onto** *module SubclassOf* | Insertion du concept dans la structure taxonomique de l'ontologie | | — | |
| | Interprétation sémantique des termes | **SSI** | | | | | | | | |
| | Construction des arbres conceptuels | — | | | | | | | | |
| **Construction des diagrammes des relations binnaires ad hoc** | Sélection des relations | — | Analyse syntaxique détaillée | *LINK-PARSER* | Identification des relations *'SubTopic of'* | **Text2Onto** *module SubtopicOf* | Identification et insertion dans l'ontologie des relations | | — | |
| | Sélection des exemples | — | Sélection des exemples | — | | | | | | |
| | Apprentissage de règles | *C4.5* | Apprentissage de règles | *Propal* | Identification des relations générales | **Text2Onto** *module Relation* | | | | |
| | Identification des relations | — | Identification des relations | — | | | | | | |
| **Construction du dictionnaire de concepts** | — | — | — | — | Identification des relations *'Instance Of'* | **Text2Onto** *module InstanceOf* | Insertion de l'instance dans l'ontologie | | — | |

TAB. 2 – *Correspondances entre outils et tâches pour chaque approche. Les outils testés sont en caractères gras.*

| | Manuel | *TermExtractor* | | *Text2Onto* | *TermRaider* | |
|---|---|---|---|---|---|---|
| | | Un seul fichier | Plusieurs fichiers | | Un seul fichier | Plusieurs fichiers |
| N° total de termes | 37 (30c+7i) | 18 | 4 | 97 (84c+13i) | 0 | 105 |
| N° de termes simples (formés d'un seul mot) | 24 (20c+4i) | 1 | 1 | 70 (58c+12i) | 0 | 74 |
| N° total de termes composés de deux mots | 9 (9c+0i) | 11 | 2 | 24 (24c+0i) | 0 | 9 |
| N° total de termes composés de plus de deux mots | 4 (1c+3i) | 6 | 1 | 3 (2c+1i) | 0 | 22 |
| N° termes communs avec le résultat manuel | – | 8 | 3 | 27 | – | 18 |
| N° termes communs / N° total de termes du résultat manuel | – | 21% | 10% | 73% | – | 49% |
| N° termes communs / N° total de termes de l'outil | – | 44% | 75% | 28% | – | 17% |
| N° de concepts parmi les termes communs | – | 5 | 2 | 21 | – | 12 |
| N° d'instances parmi les termes communs | – | 3 | 1 | 6 | – | 6 |

TAB. 3 – *Comparaison du nombre et de la longueur (en nombre des mots) des termes extraits avec les différents outils.*

### 3.2.2 Analyse des résultats des outils qui identifient des relations

Trois des outils que nous avons testés, *Text2Onto*, *SSI* et *SPRAT*, peuvent identifier des relations entre des concepts.

Il est difficile de comparer les relations identifiées manuellement (tâches 2 et 3 de Methontology) avec les relations identifiées par *Text2Onto* : *Text2Onto* a identifié 84 concepts et 13 instances alors que manuellement on obtient 30 concepts et 7 instances, avec seulement 19 concepts communs. Mais une simple analyse des relations extraites par Text2Onto peut nous aider à formuler quelques jugements quant à leur validité.

Ainsi, *Text2Onto* permet l'identification de plusieurs types de relations entre les concepts, notamment des relations avec une sémantique bien définie, comme les relations taxonomiques (nommées *'is a'*), mais aussi des relations sans sémantique précise, comme par exemple les relations *'SubTopicOf'* qui expriment une certaine relativité entre concepts. A l'aide de patrons *Text2Onto* peut également identifier des relations dont la signification est fondée sur le verbe liant deux concepts dans le texte.

*Text2Onto* propose trois algorithmes pour l'identification des relations taxonomiques. Un premier, basé sur l'utilisation de patrons, a permis l'identification de 4 relations, mais aucune de ces relations ne concerne notre domaine d'intérêt (la construction des ontologies). Le deuxième algorithme, basé sur l'inclusion des termes simples dans des termes "complexes" (technique structurelle), a identifié 26 relations. Bien que quelques unes soient superflues, elles

restent globalement valides, 5 d'entre elles se retrouvant même parmi les relations identifiées manuellement. Le troisième algorithme, qui utilise WordNet, a identifié 119 relations, dont une seule a été identifiée manuellement.

Concernant les autres relations, *TextOnto* a identifié 6935 relations *'sous topiques de'* et 4 relations à signification fondée sur des verbes. Les premières représentent toutes les combinaisons possibles entre les concepts, et donc leur validité est très discutable. Et, aucune des relations à signification fondée sur des verbes ne se retrouve parmi les relations identifiées manuellement.

Comme l'outil *SSI* n'accepte en entrée que 10 termes simples (composés d'un seul mot) au maximum, nous avons choisi de le tester sur 10 termes simples choisis parmi les termes identifiés manuellement. Le résultat obtenu est un réseau sémantique contenant 57 noeuds reliés par 5 types de relations. Dans le processus de construction du réseau sémantique l'outil a identifié de façon automatique un sens pour chaque terme. Pour deux des dix termes les sens que l'outil a retenus ne sont pas ceux concernant le domaine d'intérêt. Notons également que dans le réseau sémantique il n'y a aucun lien direct entre les termes donnés en entrée.

Comme nous l'avons précisé dans la section précédente, *SPRAT* n'a pas construit d'ontologie pour nos ressources textuelles. Dans l'ontologie construite sur les ressources plus volumineuses les 14 concepts sont reliés uniquement par des relations *'is a'*.

### 3.2.3 Discussion

Malgré la taille restreinte du corpus textuel utilisé ici pour les test, nous pouvons dégager de nos comparaisons plusieurs aspects significatifs.

Pour l'extraction des concepts nos expérimentations mettent en évidence des résultats très hétérogènes et peu de consensus entre les différents outils. Le seul indicateur vraiment pertinent pour une comparaison statistiquement significative est la proportion de termes associés aux concepts bien identifiés eu égard au résultat obtenu manuellement.

Notons que tous les outils présentés calculent pour les termes identifiés divers indices (comme leur score TFIDF, etc.). Devant l'hétérogénéité des résultats obtenus nous n'en avons pas tenu compte dans l'analyse proposée ici. Cependant, il pourra être intéressant d'analyser une éventuelle corrélation entre ces indices et la pertinence des résultats par rapport aux résultats obtenus manuellement.

Text2Onto a identifié le plus grand nombre des relations taxonomiques qui peuvent être considérées valides. Le faible nombre de ces relations qui font partie de l'ontologie construite manuellement s'explique par le fait qu'il y a seulement 19 concepts en commun entre Text2Onto (84 concepts) et l'ontologie manuelle (30 concepts). Cependant, la validité des résultats de *SPRAT* semble croître avec l'augmentation de la taille des ressources textuelles. Comme les termes ne sont pas directement liés dans le réseau sémantique qu'il construit, il nous semble difficile d'utiliser *SSI* pour l'identification de relations entre concepts, bien que ce fût l'un des buts de ses créateurs (Velardi et al., 2007).

Le passage à l'échelle, qui est un critère majeur avec l'évolution des ressources disponibles, n'est pas possible pour *GlossExtractor* et *SSI*. En revanche, les quatre autres outils ne posent pas de problèmes particuliers si ce n'est des limitations technologiques du réseau (pour les services *SPRAT* et *TermRaider*) et la disponibilité des serveurs où les outils sont hébergés.

Le dernier point de notre analyse concerne les ressources et les techniques que les divers outils emploient pour obtenir leurs résultats. OntoLearn se contente d'identifier des termes

dans les textes reçus en entrée mais fait appel à des ressources externes pour l'identification des définitions et des sens des termes, tout comme pour l'identification des relations entre concepts. Alvis utilise massivement l'analyse syntaxique des textes d'entrée mais fait appel à des experts pour la validation des résultats. Text2Onto combine l'utilisation des textes d'entrée (pour identifier des concepts, des instances et des relations) et des ressources externes (pour des relations). SPRAT combine l'utilisation des ressources externes (dictionnaires pour la validation de termes) et l'utilisation des textes reçus en entrée, mais il a la spécificité d'utiliser un grand nombre de patrons et de règles. Ces derniers dépendent fortement du type de textes en entrée et doivent être identifiés au préalable par des experts. SPRAT cherche directement dans les textes reçus en entrée le sens et les relations associées aux concepts.

En conclusion, OntoLearn, qui cherche le sens et les définitions des termes dans des ressources externes, risque d'introduire dans l'ontologie des connaissances correspondant à d'autres domaines que celui du corpus utilisé. Ce n'est pas le cas des trois autres approches, qui cherchent le sens des termes directement dans les textes d'entrée, construisant, à priori, une ontologie qui reflète plus fidèlement les connaissances du domaine du corpus utilisé.

# 4 Conclusion

Dans la première partie de cet article nous avons positionné par rapport à un référentiel issu de Methontology les tâches de quatre approches dédieés à la construction automatique d'ontologies à partir de textes. Nous avons également positionné par rapport aux différentes tâches les outils qui sont disponibles pour être testés. Dans la deuxième partie, nous avons analysé d'un point de vue technique les outils, et nous avons réalisé une comparaison experimentale entre les résultats des outils et des résultats obtenus par construction manuelle. Les résultats des outils s'avèrent prometteurs mais ils doivent être souvent raffinés ou complétés. Cette analyse nous a permis également de mettre en évidence la façon dont les quatre approches utilisent les deux types de sources de connaissances.

Ce travail est mené dans le cadre applicatif du projet ISTA3 (*Interopérabilité de 3ème génération pour les Sous Traitants de l'Aéronautique*) qui s'intéresse à l'interopérabilité des systèmes hétérogènes et vise le développement de solutions d'interopérabilité articulées autour d'une ontologie. Dans le cadre de ce projet, nous voulons construire, si possible automatiquement, une ontologie à partir de ressources textuelles. Nous avons donc réalisé cette étude afin de sélectionner un ou plusieurs outils adéquats à notre problématique.

# Références

Aime, X., F. Furst, P. Kuntz, et F. Trichet (2009). Gradients de prototypicalité appliqués à la personnalisation d'ontologies. In *Actes de la Conférence Ingénierie des Connaissances (IC2009)*, pp. 241–252.

Bourigault, D. et G. Lame (2002). Analyse distributionnelle et structuration de terminologie. application à la construction d'une ontologie documentaire du droit. *Traitement automatique des langues 43*(1), 129–150.

Cimiano, P. et J. Volker (2005). Text2onto - a framework for ontology learning and data-driven change discovery. In A. Montoyo, R. Munoz, et E. Métais (Eds.), *2nd Eur. Semantic Web Conf.*, Volume 3513, pp. 227–238.

Corcho, O., M. Fernández-López, A. G. Pérez, et A. López-Cima (2005). Building legal ontologies with methontology and webode. *LNCS 3369*, 142–157.

Fernandez, M., A. Gomez-Pérez, et N. Juristo (1997). Methontology : From ontological art towards ontological engineering. In *Proc. of the AAAI97 Spring Symposium Series on Ontological Engineering*, pp. 33–40.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowl. Acquisition 5(2)*, 199–220.

Maynard, D., A. Funk, et W. Peters (2009a). Nlp-based support for ontology lifecycle development. In *Proc. of ISWC Workshop on Collaborative Construction, Management and Linking of Ontologies*.

Maynard, D., A. Funk, et W. Peters (2009b). Sprat : a tool for automatic semantic pattern-based ontology population. In *Proc. of the Int. Conf. for Digital Libraries and the Semantic Web*.

Navigli, R. et P. Velardi (2005). Structural semantic interconnections : A knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 27*(7), 1075 – 1086.

Navigli, R., P. Velardi, et A. Gangemi (2003). Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems 18*(1), 22–31.

Nedellec, C. (2006). Semantic class learning and syntactic resources tuning. Technical report, Deliv. 6.4a for ALVIS (Superpeer semantic Search Engine ) Project.

Osborne, J., J. Flatow, M. Holko, S. Lin, W. Kibbe, L. Zhu, M. Danila, G. Feng, et R. L. Chisholm (2009). Annotating the human genome with disease ontology. *BMC Genomics 10 Supplement 1*, 63–68.

Velardi, P., A. Cucchiarelli, et M. Pétit (2007). A taxonomy learning method and its application to characterize a scientific web community. *IEEE Trans. on Knowl. and Data Eng. 19(2)*, 180–191.

Velardi, P., R. Navigli, et P. D'Amadio (2008). Mining the web to create specialized glossaries. *IEEE Intelligent Systems 23*(5), 18 – 25.

Volker, J. et E. Blomqvist (2008). Prototype for learning networked ontologies. Technical report, Deliv. 3.8.1 for NEON Project, Instit. AIFB, Univ. of Karlsruhe.

## Summary

Over the recent years, several methodologies and tools for the automatic construction of ontologies from textual resources have been proposed. In this article we first analyze four of these approaches by comparing them with a reference approach – Methontology. We choose approaches which cover all the steps of the ontology construction process. Then, for some tools associated with the previously analyzed approaches, we analyze their performances and compare their results with manually obtained results.