SHS Web of Conferences **27**, 11008 (2016) DOI: 10.1051/shsconf/20162711008
Congrès Mondial de Linguistique Française - CMLF 2016

Création semi-automatique d'un corpus annoté pour l'analyse d'opinions

Driss Sadoun
Laboratoire MoDyCo / Université Paris-Ouest Nanterre, France.
driss.sadoun@u-paris10.fr

Résumé. Nous décrivons une méthode semi-automatique pour la création d'un corpus annoté en français. Ce corpus vise à permettre l'apprentissage d'un système d'analyse d'opinions dans des textes portant sur l'évaluation d'établissements de recherche et d'enseignement supérieur. La création de ce corpus s'effectue de manière itérative. Au cours de ces itérations une ontologie, une terminologie ainsi qu'un ensemble de patrons syntaxico-sémantiques sont créés automatiquement à partir d'annotations antérieures effectuées par des experts du domaine. Ces ressources permettent par la suite de guider l'annotation automatique de nouveaux corpus. Chaque corpus annoté automatiquement est alors soumis à une nouvelle annotation manuelle des experts. Des résultats empiriques montrent que notre méthode permet d'accélérer et de faciliter le processus d'annotation. Le corpus résultat est annoté à la fois sémantiquement et syntaxiquement. Il est disponible gratuitement.

Abstract. A semi-automatic creation of an annotated corpus for opinion mining. We describe a semi-automatic method for creating an annotated corpus in French. This corpus is intended to allow learning of an opinion mining system in texts concerning the evaluation of research and higher education establishments. The creation of this corpus is done iteratively. During these iterations an ontology, a terminology and a set of syntactic patterns are automatically created from previous annotations made by domain experts. These resources allow thereafter to guide the automatic annotation of new corpus. Each automatically annotated corpus is then submitted to a further manual annotation of the experts. Empirical results show that our method accelerates and facilitates the human annotation process. The resulting corpus is annotated both semantically and syntactically and is freely available.

1 Introduction

Depuis plusieurs années, les établissements de la recherche et de l'enseignement supérieur français sont soumis à l'évaluation d'un organisme externe. Le plus souvent cette évaluation est conduite par le Haut Conseil de l'Évaluation de la Recherche et de l'Enseignement Supérieur (HCERES). Chaque année, le HCERES a pour mission de recruter et de former les experts académiques qui participeront à l'évaluation d'un ou de plusieurs établissements. Ces évaluations mènent à la production de rapports d'évaluation qui sont publiques et accessibles à tous.

Les *rapports d'évaluation d'établissements* sont des documents relativement standardisés. En effet, ces derniers couvrent sensiblement les mêmes champs d'évaluation car leur rédaction suit un référentiel établi. Ce référentiel peut se décliner en dix domaines majeurs : Formation, Gouvernance, Relations internationales, Gestion, Pilotage, Recherche, Culture scientifique et valorisation. Chaque domaine se décline ensuite en plusieurs champs. Chaque rapport d'évaluation synthétise au sein de sa conclusion les forces et faiblesses de l'établissement évalué en fonction des champs et domaines du référentiel. Le *HCERES* produit également un rapport donnant une vue globale de l'activité d'évaluation d'une année particulière. Ce type de rapport dit d'activité contient entre autres une synthèse de l'ensemble des forces et faiblesses des établissements évalués au cours d'une même année. À cet effet, chaque appréciation positive ou négative formulée dans les conclusions des rapports est classifiée manuellement selon le domaine auquel elle réfère.

Une appréciation porte sur un domaine particulier. Au sein des rapports, l'identification d'une appréciation correspond à l'identification simultanée d'un terme dénotant un domaine et d'un terme dénotant une opinion. Les phrases (1) et (2) ci-dessous contiennent respectivement une appréciation positive portant sur le domaine *Formation* et une appréciation négative portant sur le domaine *Valorisation*. Les phrases (3) et (4) quant à elles contiennent plus d'une appréciation, ce qui est assez représentatif des phrases que l'on peut trouver dans les conclusions. En effet, cela est courant et s'explique par le nombre limité (environ à cinq par type d'opinion) de phrases de synthèse dans la conclusion. Ainsi, si les phrases sont en général bien rédigées, ce phénomène d'agrégation d'appréciations peut les rendre assez longue et complexes. Dans l'ensemble de cet article, les termes dénotant un domaine apparaissent en *gras* et les termes dénotant une opinion apparaissent en *italique*.

1. une **formation doctorale** *très attractive*.
2. une **politique de valorisation de la recherche** *peu lisible*.
3. Une **présidence** *forte* mais une **gouvernance** *à revoir*.
4. Une *difficulté* de **prévision des recettes** et un *manque* d'**approche politique dans la construction du budget**.

Ce travail de classification qui porte à la fois sur l'opinion et le domaine évalué est une tâche longue, complexe et subjective pour un être humain. L'ampleur de la tâche fait qu'elle ne s'étend pas aux champs des domaines et impose qu'elle soit distribuée entre plusieurs experts académiques, ce qui a pour résultat qu'aucun expert n'a une vue globale de l'ensemble des rapports. La quantité de rapports produits n'ayant de cesse d'augmenter, il devient impératif de trouver le moyen d'accélérer et de faciliter ce travail de classification qui correspond à de l'analyse d'opinions.

Nous proposons une analyse d'opinions automatique ayant pour objectifs : 1) de réduire considérablement la subjectivité inhérente au traitement humain en le replaçant par un traitement fondé sur une connaissance commune et partagée d'experts du domaine ; 2) d'étendre la classification aux champs des domaines pour une analyse plus fine des appréciations. 3) d'étendre l'analyse aux recommandations adressées aux établissements qui sont elles aussi synthétisées dans les conclusions de rapports. De même qu'une appréciation positive ou négative, une recommandation porte sur un domaine ou un champ particulier. Par exemple la phrase (1) ci-dessous, correspond à une recommandation adressée à l'établissement par rapport à sa politique de *relations internationales*. La phrase (2) quant à elle contient plusieurs recommandations portant sur l'*offre de formation*, la *gestion financière* et les *partenariats*.

1. *Définir* une véritable **politique des relations internationales** qui permette de donner à la vocation mondiale de l'établissement une dimension concrète dans les partenariats, les formations, les stages et les échanges.

2. *Piloter* de près l'**offre de formation** : *assurer* sa **soutenabilité financière**, *impliquer plus largement* les **acteurs socioéconomiques**, *poursuivre la mise en cohérence* de l'offre à l'échelle du pôle de recherche et d'enseignement supérieur de l'établissement.

La reconnaissance des termes pertinents pour l'identification d'une appréciation repose sur une connaissance du domaine de l'évaluation des établissements de recherche et d'enseignement supérieur. Afin de capturer cette connaissance, nous proposons de créer un corpus annoté de manière semi-automatique.

L'objet de cet article est donc la création d'un corpus annoté qui constitue une première étape vers notre objectif de développement d'un système d'analyse d'opinions par apprentissage. Ce corpus est annoté sémantiquement et syntaxiquement de sorte à permettre un apprentissage fondé sur des caractéristiques sémantique et syntaxique.

Une étape préalable à un travail d'annotation est le choix des catégories d'annotation. Dans notre cadre, ces catégories correspondent aux domaines et champs d'évaluation. Le référentiel utilisé pour guider la rédaction des rapports est loin d'être exhaustif en ce qui concerne la définition des champs d'évaluation possibles. Par conséquent, une partie de notre travail consiste à identifier les catégories qui constituent le vocabulaire conceptuel de l'évaluation de la recherche et de l'enseignement supérieur. Pour définir ce vocabulaire conceptuel de manière formelle et consensuelle nous utilisons une ontologie. Dans ce cas, le référentiel existant sert d'ensemble d'amorçage pour la construction de l'ontologie. L'ontologie est construite automatiquement au fil de l'eau à partir des catégories proposées par les experts du domaine durant différentes phases d'annotation. Cette manière incrémentale de construire l'ontologie du domaine contribue à l'originalité de notre approche.

Afin de lier les termes de la langue impliqués dans la formulation d'une appréciation à leurs catégories sémantiques, nous utilisons une terminologie. La création du corpus, de l'ontologie et de la terminologie est effectuée de manière incrémentale. À chaque incrémentation l'ontologie et la terminologie sont étendues à partir des annotations produites. Ces deux ressources sont ensuite utilisées pour l'annotation semi-automatique du corpus. Le corpus annoté, l'ontologie et la terminologie construites sont disponibles gratuitement sur demande.

2 Travaux connexes

Les deux principales approches pour l'analyse d'opinions sont les approches de classifications visant à reconnaître l'orientation sémantique globale d'un texte ou d'une phrase [8, 17] et les approches *aspect-based* [15, 1] qui visent à reconnaître et à associer une orientation sémantique à un aspect ou sujet particulier. Notre ambition est d'identifier et de compter les points forts, les points faibles et les recommandations formulées sur un domaine ou un champ d'évaluation particulier. Nous nous situons donc dans la seconde approche. Indépendamment de l'approche choisie disposer d'un corpus est un facteur clé pour le développement, l'entraînement et le test d'un système d'analyse d'opinions.

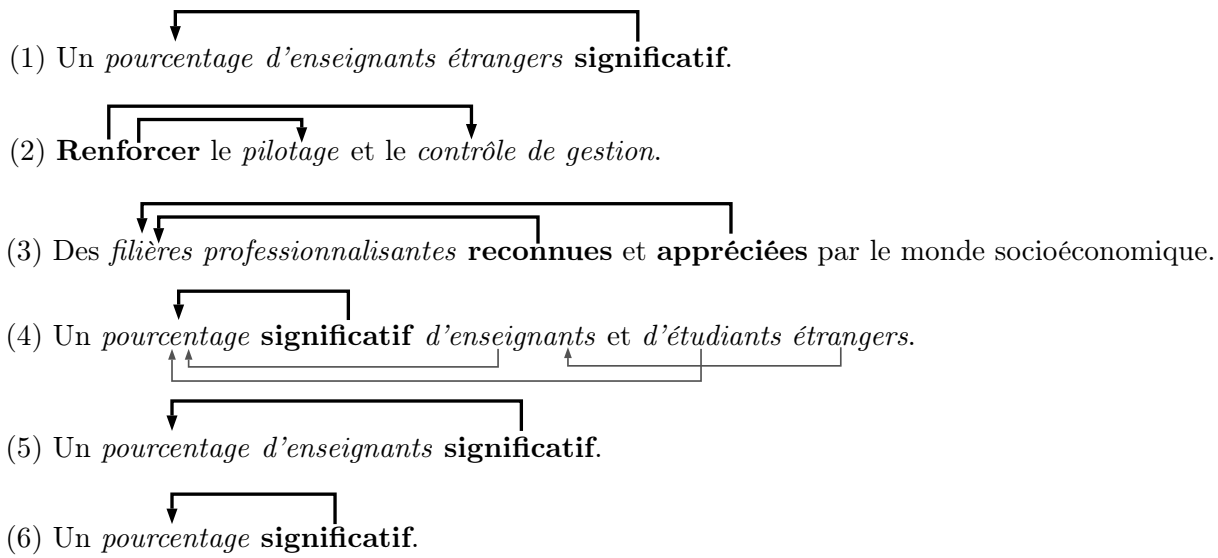
Peu de corpus annotés ont été proposés en support à l'analyse d'opinions en comparaison avec le besoin actuel. Cela, quelle que soit la langue cible. Parmi les propositions existantes de corpus annotés, on peut citer les travaux portant sur le Tchèque [23], le Norvégien [14], l'Anglais [28, 27], l'Italien [6, 20] ou le Français [7, 19]. Le petit nombre de corpus annotés existants est dû à la complexité de la tâche d'annotation manuelle qui s'avère longue et fastidieuse. Il semble alors que ce processus gagnerait à être automatisé autant que possible. Ce constat n'est pas récent, déjà dans [10] un outil d'annotation semi-automatique est proposé pour assister des utilisateurs dans l'annotation de pages *HTML*. Dans [9] un processus un peu similaire est décrit. Des annotations automatiques de textes de procédure, comme des recettes de cuisine sont soumises via une interface wiki à des utilisateurs pour être corrigés ou complétés si nécessaire. Ces deux propositions ont en commun l'utilisation d'une ontologie pour représenter les connaissances du domaine. Afin de faciliter et d'accélérer le processus d'annotation,

nous proposons donc de créer notre corpus de manière semi-automatique et guidée par une ontologie du domaine.

Par ailleurs, il a été démontré que l'utilisation de traits syntaxiques est pertinente pour l'analyse d'opinions [29, 16, 18]. En outre, des expériences ont montré que les méthodes basées sur les graphes de dépendances peuvent s'avérer nettement meilleures que les approches lexicales [14, 26]. Nous avons donc choisi d'augmenter le corpus annoté sémantiquement à partir des caractéristiques syntaxiques des mots impliqués dans les termes annotés. L'originalité de notre approche est de mêler annotations syntaxiques et sémantiques afin de permettre une annotation automatique des rapports par apprentissage. Dans la suite de cet article nous revenons sur l'intérêt pratique d'un tel choix.

3 Description du corpus à annoter

Le corpus à annoter est constitué d'appréciations issues des conclusions de 34 rapports d'évaluation d'établissements produits au cours de l'année 2013. Ces phrases sont extraites automatiquement à partir de sous-sections des conclusions décrivant les points forts et les points faibles de l'établissement ainsi que les recommandations qui lui sont adressées. En tout, ce corpus contient 692 phrases, ce qui représente environ 20 phrases par rapport. Le style rédactionnel de ces phrases n'est pas standardisé, il dépend de la prose de l'expert qui les rédige. Ces phrases peuvent être relativement longues et complexes. En effet, le nombre de mots du corpus est de 12171, ce qui signifie une moyenne de 17 mots par phrase. De plus, dans ce corpus, la majorité des termes ($\simeq 73\%$) qui dénotent un domaine ou un champ d'évaluation ou une appréciation sont des termes complexes, i.e. constitués de plusieurs mots tels que : *équipe présidentielle*, *structuration de l'offre d'enseignement* ou *manque de lisibilité*. Ces termes complexes peuvent être contigus, tels que les termes *pourcentage d'enseignants étrangers* et *contrôle de gestion* respectivement dans les phrases (1) et (2) ci-dessous ou bien être non contigus tels que les termes *pourcentage d'enseignants étrangers* et *pourcentage d'étudiants étrangers* dans la phrase (4) ci-dessous. Les phrases peuvent en outre contenir plus d'une appréciation comme il apparaît dans la phrase (2) où le terme *renforcer* qui dénote une recommandation porte sur les deux termes *pilotage* et *contrôle de gestion*. De même le terme *significatif* qui dénote une appréciation positive dans le contexte de la phrase (4) porte sur les deux termes *pourcentage d'enseignants étrangers* et *pourcentage d'étudiants étrangers*. Enfin, dans la phrase (3) le terme *filières professionnalisantes* est lié aux deux termes *reconnues* et *appréciées*. Dans les phrases ci-dessous les flèches en gras indiquent un lien entre les termes dénotant une appréciation et les termes dénotant un domaine ou champs d'évaluation. Quant aux flèches claires, elles indiquent un lien entre des mots non contigus formant un même terme complexe. Dans les deux cas, ces flèches lient les mots « tête » des termes. Parmi les mots qui forment un terme, le mot « tête » est celui qui détermine les propriétés syntaxiques du terme. Les autres mots qui entrent dans la formation du même terme peuvent être désignés comme des mots « dépendants » au sens syntaxique. Au sens sémantique ces derniers peuvent êtres considérés comme des « modifieurs ». En effet, les mots dépendants viennent préciser le sens dénotatif du *mot tête*. De plus, l'ajout ou la suppression de mots dépendants ne changent pas la distribution syntaxique entre les *mots têtes*. Par exemple, la suppression du mot *étrangers* dans la phrase (5) puis des mots *d'enseignants* dans la phrase (6) ne modifient pas la dépendance syntaxique entre le terme *significatif* et le terme pourcentage.



4 Méthode d'annotation semi-automatique

La méthode d'annotation semi-automatique que nous proposons est illustré par la figure 1. L'application de cette méthode se fait de manière incrémentale. Dans un premier temps une annotation manuelle est effectuée faisant collaborer plusieurs experts du domaine de l'évaluation d'établissements. Il en résulte une *annotation de référence* i.e. faisant consensus entre les différents annotateurs. Cette annotation de référence sert à construire les ensembles d'amorçage qui permettront l'apprentissage d'un système d'annotation automatique. Plus précisément, l'exploitation de cette annotation manuelle permet de :

1. construire une ontologie à partir des champs d'évaluation utilisés pour l'annotation du corpus.
2. construire une terminologie à partir des termes annotés en liant chaque terme au concept de l'ontologie qu'il dénote.
3. acquérir des patrons syntaxico-sémantiques caractérisant les termes annotés.

Les trois ressources ci-dessus servent alors l'annotation automatique d'un nouveau corpus. Le corpus pré-annoté automatiquement est alors soumis à une nouvelle annotation collaborative. Ce processus correspond à une itération d'annotation semi-automatique.

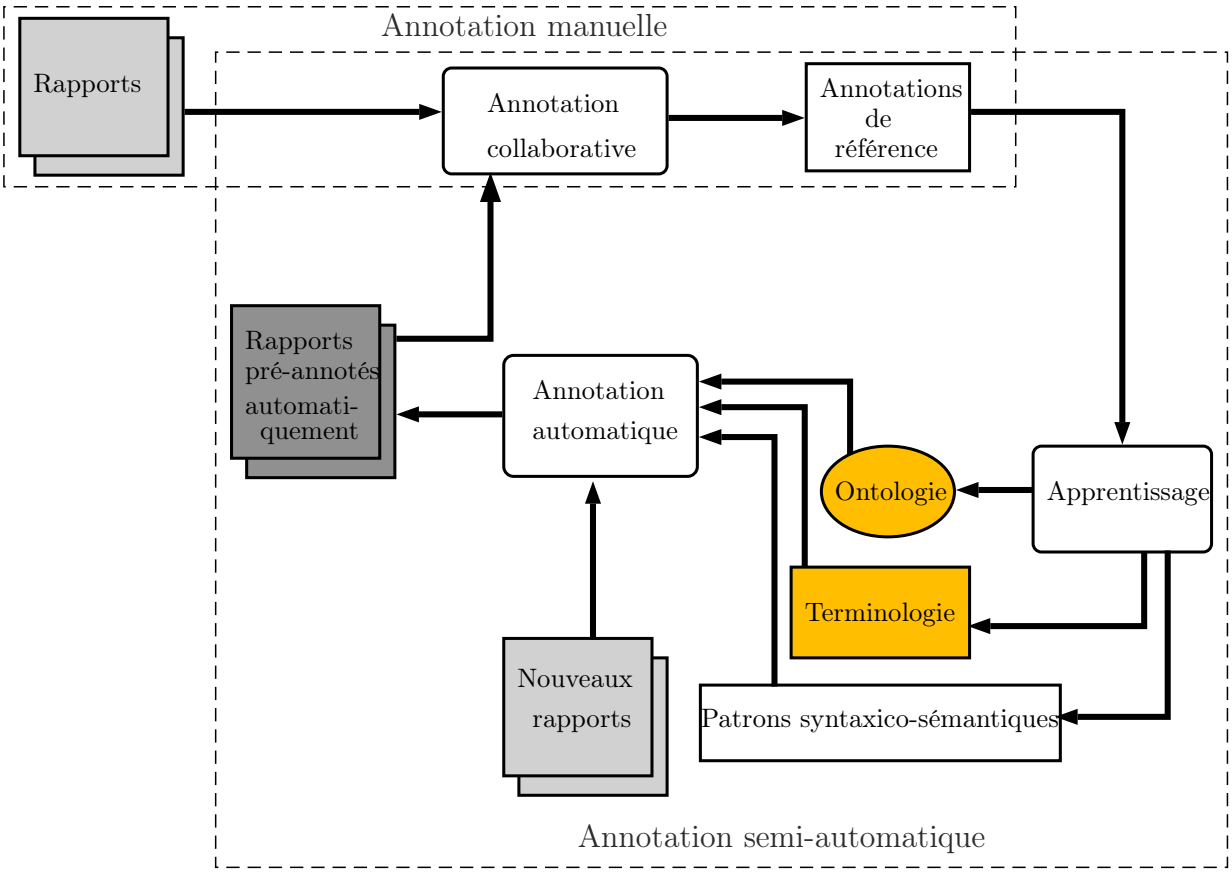


FIGURE 1 – Étapes de l'annotation semi-automatique

5 Annotation semi-automatique du corpus

Suite à une première annotation manuelle qui permet de construire un système d'annotation automatique, l'annotation semi-automatique du corpus s'effectue en deux étapes :

1. un nouveau corpus est annoté automatiquement ;
2. ce corpus annoté est soumis aux experts pour une annotation complémentaire.

Deux annotations semi-automatiques ont suivi l'annotation manuelle.

5.1 Annotation collaborative (manuelle)

L'annotation collaborative a pour objectif la construction d'une base de connaissances consensuelles. Ces connaissances sont détenues par les experts de l'évaluation d'établissements. C'est donc à ces experts qu'il revient d'annoter les corpus de conclusions de rapports. En tout, 22 experts du *HCERES* ont participé aux travaux d'annotation collaborative. Ces experts ont été répartis en six groupes d'annotateurs avec pour chaque groupe un sixième du corpus à annoter. L'annotation des conclusions s'est effectuée sous la plate-forme *Webanno* [30]. *Webanno* permet d'effectuer des annotations en ligne. Plusieurs annotateurs peuvent participer à l'annotation d'un même corpus. Chaque annotateur annote sa version du corpus sans visualiser les annotations des autres experts. Les annotations peuvent alors être comparées afin d'identifier les points d'accords et de désaccords.

Le premier challenge est de formaliser le vocabulaire conceptuel des experts, *i.e.* de construire l'ontologie en identifiant les différents champs d'évaluation appartenant aux domaines du référentiel du *HCERES*. Ce même référentiel sert d'ensemble de départ pour les catégories d'annotations, ce qui correspond aux dix domaines et à une vingtaine de champs d'évaluation. Afin d'étendre cet ensemble à l'aide de nouvelles catégories, les annotateurs ont eu la possibilité de créer de nouveaux champs pour chaque domaine lorsqu'ils estimaient que cela était nécessaire, en d'autres termes, lorsque les champs existants n'étaient pas suffisants pour caractériser les termes à annoter. Sous *Webanno*, il est possible d'attribuer à chaque catégorie d'annotation un ensemble d'étiquettes. Dans notre cadre, les catégories correspondent aux domaines d'évaluation et leurs étiquettes correspondent aux champs d'évaluation. *Webanno* autorise la création de nouvelles étiquettes durant l'annotation. Lorsqu'une nouvelle étiquette est créée, celle-ci est alors visible et utilisable par l'ensemble des annotateurs. De même, que les annotations produites par les annotateurs, les nouveaux champs proposés sont soumis à une validation consensuelle. Ils peuvent donc être validés ou rejetés, car ils reposent eux aussi sur la subjectivité de chaque annotateur.

Le protocole d'annotation proposé est assez atypique dans le sens ou l'ensemble de catégories sémantiques possibles n'est pas restreint et que chaque annotateur a pu proposer de nouvelles catégories durant son travail d'annotation. À notre connaissance, il n'existe pas de métrique pour le calcul d'accord inter-annotateur (AIA) correspondant à ce cas de figure. Pour le calcul de l'AIA, nous avons choisi de calculer la F-mesure des annotations d'un même groupe en ne tenant pas compte de l'aspect non restreint des catégories d'annotations. L'accord inter-annotateur calculé pour la première annotation manuelle s'est élevé à moins de 40%. Aussi, afin de concilier les divergences entre les annotateurs plusieurs sessions de discussion ont été organisées. Celles-ci se sont déroulées entre aux moins deux annotateurs accompagnés de l'auteur de cet article avec pour objectif d'atteindre un consensus pour chaque annotation divergente. La figure 2 correspond à un exemple de consensus faisant suite à des annotations divergentes. Dans cette figure, la première ligne correspond à l'annotation de référence, c'est à dire celle qui a fait consensus. Les quatre lignes suivantes correspondent aux annotations de quatre experts différents. On peut noter un accord total pour l'annotation du terme *développer* comme dénotant une appréciation de type *Recommandation*. Cependant, le choix du champ d'évaluation pour l'annotation du terme *sentiment d'appartenance* diffère pour chaque expert. Cela montre bien la subjectivité de la tâche d'annotation et par extension de la tâche d'identification des appréciations. On voit alors tout l'intérêt de construire un système fondé sur un ensemble de connaissances consensuelles. L'annotation retenue ici est l'étiquette *identité* qui est un champ du domaine *Gouvernance*. Les annotations de référence produites servent d'ensemble d'entraînement pour l'apprentissage d'un système d'annotation automatique.

5.2 Création de l'ontologie

La création d'une ontologie est une tâche difficile. Elle consiste à identifier les concepts et propriétés clés du domaine, ainsi que les termes qui les désignent, puis d'en fournir une définition claire, précise et non ambiguë, le tout de manière formelle et consensuelle [25, 12, 13]. Les textes sont généralement porteurs de connaissances stabilisées et partagées par des communautés de pratiques [21]. L'avancement des travaux de recherche dans le domaine du TAL, de l'extraction d'information et de l'apprentissage automatique permettent déjà l'exploitation automatique de texte pour la construction d'ontologie. Cependant, la conceptualisation d'une ontologie à partir de textes de manière complètement non-supervisée n'est pas encore envisageable. Cela tient de la nature des langues naturelles dont le sens des phrases dépend tout autant de la phrase que de son contexte d'énonciation [22] et dont le style autorise de passer sous silence certaines connaissances acceptées et partagées. Les méthodes et outils proposés sont donc le plus souvent une aide au développement permettant de réduire l'effort humain que demandent la conceptualisation d'une ontologie [4] On peut distinguer deux catégories d'outils tels que *Text2Onto* [5] ou *OntoLT* [2], Une première qui requiert l'intervention d'un humain pour valider ou rejeter les résultats d'une extraction automatique. Une seconde catégorie d'outils tels que *OntoGen* [11], *Terminae* [24] assiste l'utilisateur dans la construction d'une ontologie. L'utilisateur peut superviser l'extraction automatique d'éléments conceptuels pertinents.

Dans les deux cas, la validation n'est soumise qu'à un seul utilisateur ce qui la rend subjective. L'ontologie que nous souhaitons construire a pour objectif de représenter le vocabulaire conceptuel de l'HCERES de la manière la plus objective possible.

À chaque étape d'annotation collaborative de nouveaux champs peuvent être proposés par les annotateurs. La phase de

FIGURE 2 – Exemple d'un consensus pour une annotation divergente.

constitution d'une annotation de référence permet de ne maintenir que les nouveaux champs faisant consensus. Ces derniers viennent augmenter le vocabulaire conceptuel de l'évaluation des établissements. Afin de formaliser ce vocabulaire conceptuel, les domaines et champs d'évaluation utilisés durant les différentes annotations sont extraits automatiquement. Cette extraction correspond déjà à une structure hiérarchique conceptuel à deux niveaux, dans laquelle les champs d'évaluation sont les sous concepts des domaines d'évaluation. Cette structure est représentée sous la forme d'une ontologie OWL. Cette ontologie a pour vocation de guider l'identification des appréciations et leur catégorisation en fonction des domaine ou champs d'évaluations et des types d'appréciations qu'elles contiennent. Aussi, en plus des domaines et champs, l'ontologie représente aussi les trois types d'appréciations : *Positive*, *Négative* et *Recommandation* comme sous concepts du concept *Appréciation*. La figure 3 illustre un fragment de l'ontologie construite. On peut y voir le domaine *Appréciation* et ses trois sous domaines. Les dix domaines d'évaluation, ainsi que le détail du domaine *Gestion* contentant les cinq champs d'évaluation : *GRH*, *Gestion administrative*, *Gestion budgétaire*, *Gestion financière* et *Gestion patrimoniale*. Trois annotations successives ont permis de construire une ontologie contenant 117 concepts dont 103 champs d'évaluation. Parmi ces champs, environ 80 sont de nouveaux champs proposés et acceptés de manière consensuelle par les experts durant leur tâche d'annotation. Ces 80 nouveaux champs ont été retenus parmi plus de 100 champs proposés au fil des trois annotations.

5.3 Création de la terminologie

L'identification d'une appréciation repose en grande partie sur la reconnaissance des termes pertinents du texte et des catégories conceptuelles qu'ils dénotent. Afin d'établir le lien entre connaissances linguistiques, issues du texte et connaissances ontologiques, issues du vocabulaire conceptuel, nous proposons de construire une terminologie du vocabulaire de l'évaluation d'établissements. De manière similaire à l'ontologie, cette terminologie est extraite automatiquement à partir des annotations de référence. Cette création est faite de façon itérative à partir des termes annotés antérieurement. Les termes identifiés servent ensuite à identifier de nouveaux termes lors des phases d'annotation semi-automatique suivantes. Au sein de la terminologie, chaque terme annoté est associé à sa catégorie d'annotation *i.e.* au concept de l'ontologie qu'il dénote. Les termes identifiés sont stockés sous forme lemmatisée pour un rappel plus élevé lors de l'annotation automatique d'un nouveau corpus. L'organisation des termes au sein de la terminologie a pour objectif de permettre la distribution des appréciations en fonction de l'organisation des champs et domaines au sein de l'ontologie. Par exemple, si le terme *doctorat* est associé au concept *Formation doctorale* au sein de la terminologie et le concept *Formation doctorale* est associé au concept *Formation* au sein de

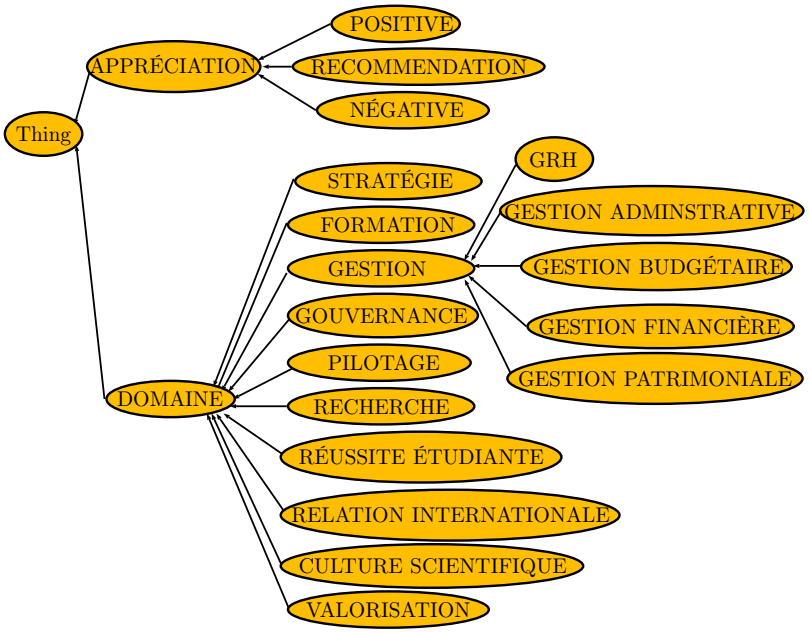


FIGURE 3 – Fragment de l'ontologie détaillant les champs du domaine *GESTION*.

l'ontologie, i.e. est un sous concept de *Formation*, alors l'appréciation identifiée comptera pour le domaine *Formation*. Si dans l'évolution de l'ontologie, le concept *Formation doctorale* devient un sous-concept du concept *Recherche*, alors l'appréciation identifiée ne comptera plus pour le domaine *Formation* et comptera alors pour le domaine *Recherche*. Ainsi, les choix conceptuels n'influencent pas le fonctionnement du système mais seulement la distribution des appréciations reconnues en fonction des domaines et champs d'évaluation. Tout au long des trois annotations successives, 1137 termes distincts ont été annotés puis créés dans la terminologie. Chacun a été associé au concept représentant un champ de l'ontologie. Lorsqu'un terme n'est pas assez précis pour être attribué à un champ il est alors associé directement à un domaine. La figure 4 illustre un exemple de termes ayant été associés au domaine *Formation* (à gauche) et de termes ayant été associés à l'appréciation *Négative* (à droite).

5.4 Apprentissage des patrons syntaxico-sémantiques

L'objectif de l'apprentissage est de capturer les contextes de formulations des termes du domaine ou d'opinion. Ces termes peuvent être complexes et contenir des mots non contigus. En outre, les mots formant ces termes peuvent être sujet à différentes

Formation

altLabel [type: XMLLiteral]
chaires annuelles

altLabel [type: XMLLiteral]
formation par la recherche

altLabel [type: XMLLiteral]
formation à l'international

altLabel [type: XMLLiteral]
mise en œuvre d'une interdisciplinarité en formation

altLabel [type: XMLLiteral]
modèle pédagogique

Négative

altLabel [type: XMLLiteral]
absence de stratégie

altLabel [type: XMLLiteral]
affaiblit

altLabel [type: XMLLiteral]
anormalement bas

altLabel [type: XMLLiteral]
carence

altLabel [type: XMLLiteral]
chaotique

FIGURE 4 – Exemple de termes pour le domaine *Formation* et l'appréciation *Négative*.

flexions de déclinaison ou de conjugaison. Une caractérisation de ces termes, doit donc en plus de tenir compte de leur catégorie sémantique, i.e. des concepts qu'ils dénotent, prendre en compte leurs traits morpho-syntaxiques. Nous proposons alors de capturer les contextes de formulation des termes annotés à l'aide de patrons syntaxico-sémantiques. Pour cela, les mots de chaque terme annoté sont analysés syntaxiquement par l'analyseur syntaxique du Français *Bonsai* [3]. Les résultats de cette analyse servent à l'apprentissage des patrons syntaxico-sémantiques. Les patrons acquis ont deux fonctions : 1) permettre de reconnaître automatiquement dans les textes les termes complexes contigus ou non contigus qui sont contenus dans la terminologie. 2) permettre de contraindre l'identification des termes en fonction de leurs caractéristiques morpho-syntaxiques. Cela pour augmenter la précision des termes reconnus. Nous distinguons deux sortes de patrons : patrons de termes simples et patrons de termes complexes.

5.4.1 Patrons de termes simples

Un terme simple est composé d'un seul mot. Dans ce cas, la reconnaissance d'un terme dépend de la présence de son lemme dans la terminologie, de sa catégorie syntaxique et de ses traits morphologiques. Par exemple, il s'avère que les termes désignant une *Recommandation* sont souvent des verbes à l'infinitif ayant pour catégories syntaxique *VINF*. Ainsi, le terme *développer* pourra être reconnu dans les textes comme dénotant une *Recommandation* contrairement aux termes *développement* ou *développé*.

5.4.2 Patrons de termes complexes

Un terme complexe est composé quant à lui de plusieurs mots. Dans ce cas, en plus de la forme lemmatisée des mots d'un terme présent dans la terminologie, de leurs catégories syntaxiques et de leurs traits morphologiques, les patrons syntaxico-sémantiques doivent capturer la manière dont les mots se combinent pour former un terme complexe. Cela revient à expliciter les relations de dépendance syntaxiques entre les différents mots d'un terme. Par exemple, le terme *sentiment d'appartenance* dénotant le champ d'évaluation *Identité*[1] est formé par les deux dépendances syntaxiques *dep* et *obj* liant respectivement le mot *sentiment* au mot *d'* et le mot *d'* au mot *appartenance* de la manière suivante : *dep(sentiment, d') - obj(d',appartenance)*. Ainsi, pour reconnaître le terme *sentiment d'appartenance* il ne suffira pas de reconnaître les mots qui le composent dans le texte mais aussi de vérifier qu'ils sont syntaxiquement liés tel que représenté par le patron syntaxico-sémantique. Par exemple, dans les phrases (1) et (2) ci-dessous, le terme *sentiment d'appartenance* sera reconnu est associé à la catégorie sémantique *Identité*, par contre il ne le sera pas pour la phrase (3).

- Un *fort* **sentiment d'appartenance** du personnel.
- Un **sentiment** *très fort* **d'appartenance** du personnel.
- Un sentiment d'insécurité et un besoin d'appartenance à un groupe social.

5.4.3 Patrons acquis

Les patrons syntaxico-sémantiques acquis couvrent en tout 116 catégories sémantiques (3 types d'appréciation, 10 domaines et 103 champs d'évaluation). Après trois annotations semi-automatiques successives, un total de 776 patrons syntaxico-sémantiques ont été acquis. Parmi eux, 728 sont des patrons de termes complexes couvrant l'ensemble des catégories sémantiques et 48 sont des patrons pour les termes simples couvrant seulement 37 catégories sémantiques. Ces deux derniers nombres montrent à quel point la proportion de termes complexes est importante dans notre corpus.

Le nombre patrons acquis (776) peut sembler important par rapports à la taille du corpus d'apprentissage (12171 mots). Cela s'explique d'abord par la distribution des patrons en 116 catégories sémantiques et leur répartition en deux types *patrons de termes simples* et *patrons de termes complexes*, mais cela tient surtout au fait que partant de zéro, chaque formulation identifiée est nouvelle et doit être acquise. Le nombre de patrons aura donc tendance à se stabiliser avec l'apprentissage de nouveaux corpus.

5.5 Annotation automatique

Comme illustré dans la figure 1, l'annotation automatique d'un corpus repose sur les connaissances acquises lors d'une annotation antérieure. L'annotation automatique est donc guidée par les connaissances de l'ontologie, de la terminologie et des patrons syntaxiques acquis. Au cours de cette annotation, chaque phrase du corpus est analysée. Chaque combinaison de

[1] Sous champ du domaine *Gouvernance*.

mots de la phrase est analysée à la recherche d'une correspondance avec l'un des termes de la terminologie. Les patrons syntaxico-sémantiques interviennent à cette étape pour reconstruire les termes pouvant appartenir à la catégorie sémantique qu'ils représentent. Les termes qui correspondent sont ensuite annotés avec le concept auquel ils sont liés au sein de la terminologie. Le corpus automatiquement annoté est ensuite soumis aux experts pour une nouvelle annotation collaborative. Les annotateurs peuvent alors valider, corriger ou ajouter des annotations manquantes.

Suite aux deux annotations semi-automatiques qui ont été effectuées, on peut distinguer trois types d'annotations au sein de l'annotation de référence produite :

- les annotations ajoutées qui correspondent aux nouvelles annotations proposées par les annotateurs. Elles représentent environ 43% des annotations de références.
- les annotations validées qui correspondent aux annotations automatiques conservées par les experts. Elles représentent environ 32% des annotations de références.
- les annotations corrigées qui correspondent aux annotations des experts contenant au moins une annotation générée automatiquement. Par exemple, le terme annoté *pilotage de la formation* contient les deux termes annotés automatiquement *pilotage* et *formation*. De même le terme *manquant de cohérence* contient le terme *cohérence* annoté automatiquement. Ces annotations représentent 25% des annotations de références.

Parmi les annotations de référence un tiers ($\simeq$32%) correspond aux annotations automatiques et plus de la moitié ($\simeq$57%) est issu des annotations automatiques. En outre, les experts annotateurs ont constaté que la pré-annotation automatique a considérablement aidé et facilité leur travail d'annotation. Ces résultats montrent que la pré-annotation automatique constitue une aide précieuse pour les annotateurs. De plus, l'annotation automatique reflète un accord consensuel entre les experts annotateurs, rendant ainsi la tâche moins subjective.

6 Caractéristiques du corpus annoté

Le corpus résultat du processus d'annotation semi-automatique contient 1792 annotations dont 932 sont attribuées à termes dénotant un domaine ou un champ d'évaluation et 860 sont attribuées à des termes dénotant une appréciation. Ces résultats sont donnés dans le tableau 1.

Type d'annotation		Nombre d'annotations
Domaine & champ		932
Appréciation		
Appréciation positive	300	860
Appréciation négative	238	
Recommandation	322	
Total		1792

TABLE 1 – Résultats de l'annotation collaborative.

En plus de l'annotation sémantique, nous avons choisi d'annoter notre corpus syntaxiquement. En effet, l'usage de dépendances syntaxiques permet de reconnaître des termes complexes de la terminologie même lorsqu'ils contiennent des mots non contigus dans les textes. En outre, la vérification des traits morpho-syntaxiques permet d'identifier les termes de manière précise.

L'ajout des traits syntaxiques aux annotations sémantiques du corpus s'effectue de manière automatique suite à chaque itération d'annotation collaborative. Chaque phrase annotée sémantiquement est analysée syntaxiquement par l'analyseur syntaxique *Bonsai* [3]. Les traits morpho-syntaxiques des mots d'une phrase sont alors ajoutés à ses annotations sémantiques. La figure 5 illustre la combinaison de traits syntaxiques et d'annotations sémantiques de la phrase : *Un sentiment très fort d'appartenance*. Les huit premières colonnes contiennent les annotations syntaxiques et les deux dernières colonnes les annotations sémantiques. Les annotations syntaxiques correspondent dans l'ordre à l'index du mot (*ID*), la forme du mot dans le texte (*FORM*), le lemme du mot (*LEMMA*), la partie du discours simplifiée du mot (*CPOS*), la partie du discours du mot (*POS*), les traits morphologiques (*FEAT*), le numéro du *nœud tête* du nœud courant (*HEAD*) qui est soit un ID ou zero (si le nœud tête est *ROOT*) et la relation de dépendance du nœud courant avec son *nœud tête* (*DEP*). En ce qui concerne les annotations sémantiques, la neuvième colonne contient les caractéristiques de l'annotation. Par exemple, la notation *IDENTITÉ[2, 5, 6] (GOUVERNANCE)* portant sur le mot numéro 2 signifie que le mot *sentiment* fait parti du terme (*sentiment d'appartenance*) contenant par ailleurs les mots 5 et 6 respectivement *d'* et *appartenance*. Ce terme dont le mot numéro 2 correspond à la *tête* a pour annotation *IDENTITÉ* qui est un champ du domaine *GOUVERNANCE*. La dixième colonne est utilisée pour les termes

complexes non contigus. Elle contient l'ID du mot qui correspond à la plus proche partie antérieure du terme non contigu. Par exemple, la notation *2]SUITE_GOUVERNANCE* associée au mot numéro 5 (*d'*) signifie que ce mot est lié au mot numéro 2 (*sentiment*) au sein d'une annotation de type *GOUVERNANCE*.

ID	FORM	LEMMA	CPOS	POS	FEATURES	HEAD	DEP	ANNOTATION	ANNOTATION LINK
1	Un	un	D	DET	g=m|n=s|s=ind	2	det	<>	<>
2	sentiment	sentiment	N	NC	g=m|n=s|s=c	0	root	<Identité[2, 5, 6] (Gouvernance)>	<>
3	très	très	A	ADV	_	4	mod	< >	<>
4	fort	fort	A	ADJ	g=m|n=s|s=qual	2	mod	<Positive[3, 4] (Appréciation)>	<>
5	d'	de	P	P	_	2	dep	< >	< 2]Suite_Gouvernance >
6	appartenance	appartenance	N	NC	g=f|n=s|s=c	5	obj	< >	< >

FIGURE 5 – Phrase annotée syntaxiquement et sémantiquement.

Nous avons choisi de représenter les annotations sur les mots de tête car dans les graphes de dépendances syntaxiques les dépendances entre les termes (simples ou complexes) désignant un domaine et ceux désignant une appréciation sont exprimés sur leurs *nœuds tête*. Par exemple, La relation de dépendance *mod* entre le nœud 4 (*fort*) et son *nœud tête* 2 (*sentiment*) signifiant que le terme *très fort* modifie le terme *sentiment d'appartenance* est exprimée entre les mots têtes *fort* et *sentiment*. L'avantage du recours aux dépendances syntaxiques réside dans le fait que quel que soit l'ordre des mots dans la phrase, que les termes soient contigus ou non, ou que des mots s'ajoutent aux *mots têtes* pour créer des termes complexes, les dépendances entre les nœuds de tête restent inchangées. Cette propriété s'avère très utile dans notre contexte ou près de 15% des termes complexes sont non contigus.

7 Discussion

L'approche d'annotation semi-automatique présentée repose sur l'utilisation d'une ontologie de domaine et d'une terminologie qui lui est associée. L'usage de ces deux ressources ne conditionne en rien la portabilité de notre approche. En effet, l'existence d'une ontologie de domaine et d'une terminologie servirait l'amorce d'une annotation semi-automatique. L'itération des annotations permettrait ensuite d'augmenter le contenu de chacune de ces deux ressources. Dans le cas ou ces ressources seraient inexistantes, elles seraient alors construites à partir de la première annotation manuelle et étendues à chaque itération.

L'annotation du corpus est effectuée sur deux niveaux. Un premier sémantique, dans lequel les termes de chaque phrase sont annotés par un champ ou domaine d'évaluation ou par un type d'opinion (positif, négatif ou recommandation). Un second syntaxique qui à chaque mot associe ses caractéristiques morpho-syntaxiques. Ces deux niveaux se complètent sans être inter-dépendants. Les annotations manuelles ne dépendent pas de l'analyse syntaxique de même que les résultats de l'analyse syntaxique ne dépendent pas l'annotation sémantique. La combinaison des deux niveaux sert à l'acquisition de patrons syntaxico-sémantiques qui servent à l'annotation automatique de nouveaux corpus. Quant à l'apprentissage du système d'analyse d'opinions elle pourra se faire sur les annotations du niveau sémantique ou de la combinaison des deux niveaux.

L'usage d'un niveau syntaxique sert notamment à l'identification de termes négatifs qui correspondent souvent à des termes complexes contenant des mots sous la portée d'une négation comme par exemples : *peu lisible*, *non maîtrisée* ou *mal intégrés*. La portée de la négation se traduit en général par une dépendance syntaxique faisant le lien entre le mot qui porte la négation et le mot sur lequel porte la négation. Cela permet de reconnaître des termes négatifs tels que *pas clair* ou *pas à la hauteur* même lorsque les mots qui les composent ne sont pas contigus comme par exemple dans *pas* assez *clair* ou *pas* vraiment *à la hauteur*.

Nous avons choisi d'annoter les recommandations car elles portent elles aussi sur les champs et domaines d'évaluation. L'objectif étant d'identifier le maximum de termes du domaine et les lier aux concepts qu'ils dénotent. Par ailleurs, la formulation d'une *recommandation* peut suggérer une opinion neutre, positive ou négative. Par exemple, la phrase (1) ci-dessous est de polarité neutre car elle n'est pas explicitement fondée sur un point fort ou un point faible de l'établissement, contrairement à la recommandation de la phrase (2) qui explicite un point positif qu'il faut maintenir ou des recommandations des phrases (3) et (4) qui suggèrent des améliorations à apporter pour pallier des points négatifs constatés par les experts. La classification des recommandations en fonction de leur polarité fait partie de nos perspectives.

1. Veiller à ce que le CA joue son rôle décisionnel en s'appuyant sur les nouvelles instances de gouvernance.
2. Maintenir l'exemplarité de la formation à distance.
3. Améliorer l'accueil et l'intégration de tous les étudiants, les suivre après leur diplôme et organiser un réseau des diplômés.

4. Clarifier l'organisation interne, notamment les responsabilités respectives entre politiques et administratifs pour le pilotage de l'établissement.

8 Conclusion

Dans cet article, nous avons présenté une méthode semi-automatique pour la création d'un corpus annoté. Ce corpus vise à permettre l'apprentissage d'un système d'analyse d'opinions dans le domaine de l'évaluation des établissements de recherche et d'enseignement supérieur. La méthode proposée est incrémentale, elle permet de créer ou d'étendre à chaque nouvelle itération les ressources suivantes : une ontologie formalisant les connaissances consensuelles d'experts en matière d'évaluation d'établissement de recherche et d'enseignement supérieur ; une terminologie faisant le lien entre les termes identifiés (annotés) dans les rapports d'évaluation et les concepts de l'ontologie ; un ensemble de patrons syntaxico-sémantiques caractérisant la formulation de termes dénotant des domaines ou champs d'évaluation ou des appréciations. Ces ressources sont alors exploitées pour effectuer l'annotation automatique de nouveaux corpus qui sont ensuite soumis aux experts pour une nouvelle itération d'annotation collaborative. La pré-annotation automatique qui sert à chaque itération se révèle être d'une aide précieuse pour les annotateurs humains. La méthode que nous avons proposé se veut indépendante du domaine. Elle prend en compte les catégories utilisées pour l'annotation d'un texte pour construire une ontologie, puis extrait les termes annotés afin de les associer aux concepts qu'ils dénotent au sein d'une terminologie. Le corpus que nous avons présenté est en français, cependant nous pensons que cette approche pourrait s'adapter à d'autres langues tant qu'un corpus de domaine est disponible et qu'un analyseur syntaxique existe pour la langue cible.

Références

[1] BAGHERI, A., SARAEE, M., AND DE JONG, F. ADM-LDA : An aspect detection model based on topic modelling using the structure of review sentences. *Journal of Information Science 40*, 5 (2014), 621–636.

[2] BUITELAAR, P., OLEJNIK, D., AND SINTEK, M. A protégé plug-in for ontology extraction from text based on linguistic analysis. In *The Semantic Web : Research and Applications* (2004), vol. 3053, pp. 31–44.

[3] CANDITO, M., CRABBÉ, B., AND DENIS, P. Statistical French Dependency Parsing : Treebank Conversion and First Results. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (2010).

[4] CIMIANO, P., MÄDCHE, A., STAAB, S., AND VÖLKER, J. Ontology learning. In *Handbook on Ontologies*. 2009, pp. 245–267.

[5] CIMIANO, P., AND VÖLKER, J. Text2onto - a framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)* (2005), vol. 3513, pp. 227–238.

[6] CROCE, D., GARZOLI, F., MONTESI, M., CAO, D. D., AND BASILI, R. Enabling Advanced Business Intelligence in Divino. In *Proceedings of the 7th International Workshop on Information Filtering and Retrieval co-located with the 13th Conference of the Italian Association for Artificial Intelligence* (2013), pp. 61–72.

[7] DAILLE, B., DUBREIL, E., MONCEAUX, L., AND VERNIER, M. Annotating opinion-evaluation of blogs : the Blogoscopy corpus. *Language Resources and Evaluation 45*, 4 (2011), 409–437.

[8] DALAL, M. K., AND ZAVERI, M. A. Semisupervised Learning Based Opinion Summarization and Classification for Online Product Reviews. *Appl. Comp. Intell. Soft Comput.* (2013).

[9] DUFOUR-LUSSIER, V., BER, F. L., LIEBER, J., MEILENDER, T., AND NAUER, E. Semi-automatic annotation process for procedural texts : An application on cooking recipes. *CoRR* (2012).

[10] ERDMANN, M., MAEDCHE, A., SCHNURR, H.-P., AND STAAB, S. From manual to semi-automatic semantic annotation : About ontology-based text annotation tools. In *Proceedings of the COLING - Workshop on Semantic Annotation and Intelligent Content* (2000).

[11] FORTUNA, B., GROBELNIK, M., AND MLADENIC, D. Ontogen : semi-automatic ontology editor. In *Proceedings of the 2007 conference on Human interface : Part II* (2007), pp. 309–318.

[12] GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing ? *International Journal of Human-Computer Studies 43*, 5–6 (1995), 907 – 928.

[13] GUARINO, N. Some ontological principles for designing upper level lexical resources. In *Proceedings of the First International Conference on Language Resources and Evaluation* (1998).

[14] HAMMER, H. L., SOLBERG, P. E., AND ØVRELID, L. Sentiment classification of online political discussions : a comparison of a word-based and dependency-based method. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (2014), pp. 90–96.

[15] HU, M., AND LIU, B. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2004), pp. 168–177.

[16] JIANG, L., YU, M., ZHOU, M., LIU, X., AND ZHAO, T. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1* (2011), pp. 151–160.

[17] KHAN, F. H., BASHIR, S., AND QAMAR, U. TOM : Twitter Opinion Mining Framework Using Hybrid Classification Scheme. *Decis. Support Syst. 57* (2014), 245–257.

[18] LAPPONI, E., READ, J., AND OVRELID, L. Representing and Resolving Negation for Sentiment Analysis. In *Data Mining Workshops (ICDMW)* (2012).

[19] LARK, J., MORIN, E., AND PEÑA SALDARRIAGA, S. CANÉPHORE : a French corpus for aspect-based sentiment analysis evaluation. TALN, 2015.

[20] MELE, F., SORGENTE, A., AND VETTIGLI, G. An Italian Corpus for Aspect Based Sentiment Analysis of Movie Reviews. In *First Italian Conference on Computational Linguistics CLiC-it* (2014).

[21] MONDARY, T., DESPRÉS, S., NAZARENKO, A., AND SZULMAN, S. Construction d'ontologies à partir de textes : la phase de conceptualisation. In *Actes des 19èmes Journées Francophones d'Ingénierie des Connaissances (IC'08)* (2008), pp. 87–98.

[22] RUSSELL, S. J., NORVIG, P., CANDY, J. F., MALIK, J. M., AND EDWARDS, D. D. *Artificial intelligence : a modern approach*. Prentice-Hall, Inc., 1996.

[23] STEINBERGER, J., BRYCHCÍN, T., AND KONKOL, M. Aspect-Level Sentiment Analysis in Czech. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (2014).

[24] SZULMAN, S. Une nouvelle version de l'outil terminae de construction de ressources termino-ontologiques. In *22èmes journées francophones d'Ingénierie des Connaissances (poster)* (2011), p. 3.

[25] USCHOLD, M., AND KING, M. Towards a methodology for building ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95* (1995).

[26] VILARES, D., ALONSO, M. A., AND GÓMEZ-RODRÍGUEZ, C. On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages. *Journal of the Association for Information Science and Technology* (2015).

[27] WACHSMUTH, H., TRENKMANN, M., STEIN, B., ENGELS, G., AND PALAKARSKA, T. A Review Corpus for Argumentation Analysis. In *Computational Linguistics and Intelligent Text Processing*, vol. 8404. 2014, pp. 115–127.

[28] WIEBE, J., WILSON, T., AND CARDIE, C. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation 39*, 2-3 (2005), 165–210.

[29] WU, Y., ZHANG, Q., HUANG, X., AND WU, L. Phrase Dependency Parsing for Opinion Mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 3* (2009), pp. 1533–1541.

[30] YIMAM, S. M., DE CASTILHO, R. E., GUREVYCH, I., AND BIEMANN, C. Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations* (2014), pp. 91–96.