Extraction d'indicateurs de construction collective de connaissances dans la formation en ligne

Alexandre Baudrillart[1, 2]
(1) Université Stendhal Grenoble 3, BP 25, 38040 Grenoble Cedex 9, France
(2) Université de Lyon, CNRS INSA-Lyon, UMR5205 F-69621, France
alexandre.baudrillart@u-grenoble3.fr

RÉSUMÉ

Dans le cadre d'apprentissages humains assistés par des environnements informatiques, les techniques de TAL ne sont que rarement employées ou restreintes à des tâches ou des domaines spécifiques comme l'ALAO (Apprentissage de la Langue Assisté par Ordinateur) où elles sont omniprésentes mais ne concernent que certaines dimensions du TAL. Nous cherchons à explorer les possibilités ou les performances des techniques voire des méthodes de TAL pour des systèmes moins spécifiques dès lors qu'une dimension de réseau et de collectivité est présente. Plus particulièrement, notre objectif est d'obtenir des indicateurs sur la construction collective de connaissances, et ses modalités. Ce papier présente la problématique de notre thèse, son contexte, nos motivations ainsi que nos premières réflexions.

ABSTRACT

Collaborative Knowledge Building Indicators Extraction in Distance Learning

Natural Language Processing techniques are still not very much used within the field of Technology Enhanced Learning. They are restricted to specific tasks or domains such as CALL (standing for Computer Assisted Language Learning) in which they are ubiquitous but do not match every linguistic aspect they could process. We are seeking to explore possibilities or performances of thoses techniques for less specific systems including a network or community aspect. More precisely, our goal is to get indicators about collective knowledge building and its modalities. This paper presents the problem and the background of our thesis problem, as well as our motivation and our first reflections.

MOTS-CLÉS : TAL, EIAH, formation en ligne, socio-constructivisme, acquisition des connaissances, apprentissage collaboratif en ligne.

KEYWORDS: NLP, TEL, distance learning, socio-constructivism, knowledge aquisition, collaboration, CSCL.

Introduction

Nous abordons dans cet article une problématique liée à un potentiel apport des traitements automatiques de la langue aux Environnements Informatiques pour l'Apprentissage Humain (EIAH) avec une finalité d'amélioration des performances pédagogiques.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 337–350,
Grenoble, 4 au 8 juin 2012.

Les Environnements Informatiques pour l'Apprentissage Humain permettent à des apprenants, à des tuteurs ou à des professeurs d'interagir avec et au travers d'un système. Les apprenants vont grâce à lui pouvoir prendre part à des situations d'apprentissages, participer individuellement ou en groupe à des activités. Le rôle du système est de faciliter l'apprentissage. Une manière de favoriser la réussite d'un apprenant dans son entreprise peut être de guider l'apprenant et de lui fournir des retours (*feedback*), sur les performances d'apprentissage. Ces retours peuvent présenter des informations plus ou moins brutes, ou des conseils précis ou directifs. Ce sont des indicateurs. Afin qu'un apprenant puisse en tirer un bénéfice, ces retours doivent être interprétables et exploitables.

Dans le cadre d'apprentissages collectifs à distance, un des rôles importants de l'environnement est de permettre l'interaction et la discussion entre les différents protagonistes et principalement entre les apprenants. C'est grâce à ce dispositif qu'un partage d'informations peut avoir lieu et que des débats peuvent s'instaurer, permettant ainsi la construction collective de connaissances.

Une des fonctionnalités qu'offre ce genre de système est la conservation de traces d'activités. Ces traces sont donc de deux types : des traces d'interactions avec le système, c'est-à-dire les actions réalisées par les utilisateurs et médiatisées par l'Interface Homme-Machine, et des traces d'interactions entre agents humains, supportées par la langue et donc textuelles, au travers du système (dans un cadre collaboratif). Ces échanges textuels véhiculent des idées, des états d'esprit ou des comportements sociaux qui participent à une construction individuelle et collective de connaissances. Les traces textuelles et d'actions s'organisent sur un axe temporel ce qui permet d'accéder à une dimension séquentielle de l'activité des participants. À ces traces s'ajoute la conservation de documents produits individuellement ou en collaboration qui sont alors le résultat attendu des échanges, qui synthétisent les idées et les cristallisent. Une production collective peut être construite en parallèle des échanges ou *a posteriori*, et est susceptible d'être augmentée, réduite ou refondue, à plusieurs reprises. Elle exprime un état des connaissances collectives.

La problématique que nous voulons aborder consiste à évaluer les possibilités d'applications des traitements automatiques des langues actuels pour extraire des indicateurs sur le niveau, la qualité et les modalités de connaissances construites collectivement et en réseau, dès lors que des traces d'interactions en langue naturelle entre les différents apprenants ou de productions collectives sont conservées et collectées. Cette extraction, qui doit résulter de traitements automatiques sur les différentes traces accumulées, a pour but d'avoir un impact sur les performances d'apprentissage. Cette problématique est encore vague et sujet à discussion, tant en ce qui concerne les situations d'apprentissage précises concernées, le matériau textuel susceptible d'être analyser et les indicateurs à produire.

Cet article a pour but de présenter les motivations, apporter des justifications de notre problématique. Il s'organise de la façon suivante. Nous commencerons par exposer un modèle de notation constituant pour nous une forme canonique de ce que nous pourrions vouloir automatiser : une première motivation directe. Nous présenterons ensuite dans quelle mesure sont utilisées certaines méthodes de TAL pour les EIAH et notamment pour les systèmes ALAO pour éclairer non pas un verrou scientifique à lever, mais une pluridisciplinarité encore embryonnaire : apportant des motivations indirectes. Nous exposerons ensuite en quoi l'Analyse Sémantique Latente, convient pour représenter la connaissance, pour simuler un modèle d'apprentissage ou pour évaluer des notions acquises par l'apprenant en tant qu'individu isolé. Nous essayerons de montrer qu'il est possible d'aller d'une approche statistique et connexionniste pour l'individu à un

modèle plus symbolique et dialogique pour un groupe d'apprenants, et donc à caractère social, ce qui rejoint notre problématique. Nous terminerons sur la nécessité d'amorcer nos recherches en nous focalisant sur un sous-problème et en construisant un corpus d'observation.

1 Motivation directe

1.1 Présentation des travaux

Dans (Hou et Wu, 2011), sont étudiées les caractéristiques des discussions synchrones médiatisées à but pédagogique. L'intérêt se porte sur l'impact d'une discussion synchrone et de sa médiatisation informatique par messagerie instantanée sur la construction collective de connaissances, les types d'interactions et les degrés de coordination. Un autre objectif est de voir émerger des motifs séquentiels de comportements c'est-à-dire de types de message caractérisant une discussion de haute qualité ou une discussion de basse qualité. Pour cela, une expérience à deux facettes est mise en place. Cette étude requiert l'intervention d'experts humains pour catégoriser et évaluer les interactions textuelles individuellement d'une part et la qualité des discussions d'autre part.

1.2 L'expérience

L'expérience consiste à observer à long terme, trois mois, des groupes d'étudiants de moins de dix membres chacun, constituant au total quarante apprenants, discuter et débattre par le biais d'une application de messagerie instantanée. Ces discussions portent sur un sujet fixé par leur professeur ; elles doivent mener les étudiants à des conclusions et des synthèses.

La première facette de l'expérience est une analyse quantitative des messages. Des experts humains, possédant des compétences psychologiques, classent les messages en types d'interaction représentant un certain comportement de l'apprenant au sein de chacune des discussions, comportements eux-mêmes regroupés en quatre catégories : élaboration de la connaissance, organisation et collaboration entre apprenants, interactions sociales et hors-sujet. Puis, ce sont les qualités des discussions entières qui sont évaluées par l'enseignant, expert dans la discipline qu'il enseigne et sur lesquelles portent les discussions. Elles sont notées selon quatre critères : clarification du sujet, collecte d'informations, profondeur d'analyse et conclusion. Ces notes permettent ensuite de construire deux catégories : haute qualité et basse qualité. Peu de précisions sont données à ce sujet.

La deuxième facette de l'expérience concerne les suites de comportements. Afin de faire émerger des continuités, des discontinuités ainsi que des dépendances entre comportements ou catégories de comportement, une analyse séquentielle des suites de comportements statistiquement significatives est réalisée.

1.3 Résultats

Les résultats quantitatifs concernant toutes les discussions montrent qu'entre la moitié et deux tiers des contributions sont hors-sujet.

Une autre part de 30 % est principalement constitué de messages concernant des échanges sur le savoir académique. Le reste concerne la coordination des étudiants ou des messages à caractères sociaux comme des remerciements ou des encouragements.

Les résultats relatifs à chaque qualité de discussion apparaissent et discriminent haute qualité et basse qualité. Le premier, le plus flagrant, met en évidence que les discussions de haute qualité ont généré quatre fois plus de contributions, ce qui souligne un rythme soutenu et une profondeur des discussions, ainsi qu'une motivation de la part des apprenants. Ces discussions sont plus variées dans l'élaboration de la connaissance, et l'identification de désaccords et la négociation du sens sont plus présentes. La dimension organisationnelle est presque inexistante au sein des discussions de basse qualité alors qu'elle apparaît de manière pertinente dans celles de haute qualité indiquant que ces apprenants explicitent la coordination de leurs démarches. Selon les auteurs, il résulte de la comparaison des chiffres que des interactions sociales telles que des encouragements ou des félicitations sont un ciment entre la construction de la connaissances et la coordination. Une autre observation intéressante est l'apparente indépendance des contributions hors-sujet vis-à-vis de la qualité des discussions. Ces discussions peuvent être à l'origine d'un climat propice à une meilleure qualité de discussion.

L'analyse séquentielle permet de mettre en évidence qu'il n'y a pas de véritable continuité d'un comportement particulier de construction de connaissance. Par contre, il existe des suites de comportements variés de construction de la connaissance, garante du maintien du focus des apprenants sur le sujet et de discussions plus approfondies. En outre, ces motifs séquentiels significatifs ne traduisent aucune dépendance des contributions hors-sujet vis-à-vis des autres catégories.

1.4 Discussion sur une automatisation

Nous pouvons nous interroger quant à la capacité de techniques informatiques à automatiser ces protocoles expérimentaux et à aboutir des conclusions similaires. Il faut noter que chaque analyse, classification ou évaluation est réalisée par des experts humains et qu'aucune remarque n'est donnée au sujet de traitements pouvant automatiser ces processus. La classification des messages est effectuée par des experts en psychologie, et les types de comportements sont eux-mêmes regroupés en catégories selon, deux niveaux. La qualité des discussions est évaluée, selon un barème par un expert du domaine : l'enseignant.

Identifier automatiquement la catégorie d'un message pourrait permettre la détection de séquences particulières et émettre des hypothèses sur la direction que prennent les échanges permettant d'inférer des conseils, des pistes ou encore simuler un participant virtuel (proche d'un *tuteur intelligent*), à cette discussion afin d'améliorer discussion et apprentissage par son biais. Ces processus décisionnels ne sont pas triviaux. Cette perspective nous amène à nous interroger sur la pluridisciplinarité qui existe entre EIAH et TAL : a-t-on les moyens de répondre à cette automatisation les méthodes voire les techniques et les ressources dont dispose la communauté ?

2 Motivations indirectes

D'autres motivations à notre problématique proviennent d'une utilisation très modérée et non entière du TAL. Pourtant, les connaissances ou les échanges à traiter au sein d'environnement d'apprentissage sont supportés par un matériau langagier.

Nous allons exposer des situations d'apprentissages assistées par ordinateur afin de montrer qu'il y a encore beaucoup d'opportunités à marier EIAH et TAL par l'intermédiaire d'un état de l'art encore partiel, en nous appuyant sur une classification proposée dans (Gurevych *et al.*, 2009).

L'actuelle utilisation des technologies du TAL au sein du champ des EIAH y est dépeinte et fractionnée en quatre catégories : génération automatique d'exercices, évaluation automatique de dissertation, aide à la lecture et à l'écriture et gestion de contenus et apprentissages collaboratif(s). La suite détaille à quelles activités pédagogiques ou éléments de l'ingénierie pédagogique fait référence chacune de ces catégories, et présenter en quoi est employé le TAL.

2.1 Génération automatique d'exercices

La première catégorie regroupe entre autres la génération automatique d'éléments d'exercices tels que les questions à choix multiples (Karamanis, 2006), ou d'exercices entiers tels que les exercices lacunaires (Lee et Seneff, 2007). Elle réunit la construction des intitulés, des réponses associées, la notation automatique des résultats ainsi que l'évaluation de l'efficacité de ces tests pour juger la qualité d'acquisition de connaissances et pour différencier les « bons » étudiants des « moins bons» .

La démarche générale consiste à extraire automatiquement de textes, à l'aide de motifs syntaxiques, la réponse à une question, notamment au sein de phrases définitoires. Pour les QCM, il faut ensuite générer l'énoncé qui va amener à la bonne réponse, en transformant affirmations en interrogations, en inversant la construction syntaxique pour l'anglais et en choisissant un des fameux pronoms 'WH'. En revanche, pour de simples exercices lacunaires, il suffit de capturer le matériau authentique qui constitue en lui-même l'objet du test. Il ne reste alors plus qu'à ôter l'élément qui constituera la lacune, et d'indiquer l'attente qui est souvent générique.

Dès lors qu'un choix multiple est envisagé, une dernière étape intervient dans la génération de l'exercice : le choix de « *distractors*» . Ces éléments représentent les autres choix possibles que les réponses attendues dont leur rôle et leur choix répondent à des critères particuliers comme une proximité sémantique suffisante sans pour autant installer une ambiguïté. Ces exercices attendent des réponses fermées, simplifiant ainsi la vérification de celles proposées par les étudiants : l'évaluation automatique reste donc relativement aisée.

Il faut noter que ces types d'exercices portent souvent sur des questions de vocabulaire, d'orthographe ou de conjugaison et ne concernent donc que l'apprentissage de la langue elle-même, maternelle ou seconde. Près de 90% des références de cette catégorie citées dans (Gurevych *et al.*, 2009) concernent l'apprentissage de la langue. Leurs emplois ne se limitent pas à cet apprentissage mais peuvent aborder d'autres disciplines, pour des questions de terminologie ou de compréhension.

2.2 Évaluation automatique de dissertation

L'évaluation de résumés ou de dissertations produits par des apprenants est une tâche pouvant éprouver le TAL. Cette seconde catégorie évalue des critères que nous regroupons selon quatre principaux points : lisibilité, focalisation sur le sujet à traiter (en évitant les hors-sujet) et les thèmes ou notions abordés, qualité d'argumentation et validité des propos représentant la réelle compréhension.

L'évaluation de la lisibilité est depuis longtemps traitée et utilise notamment des méthodes numériques qui consistent par exemple à compter des descripteurs de surface. Ces derniers sont par exemple le nombre moyen de syllabes par mot, de mots ou de syntagmes par phrase, ou encore le nombre d'hapax, de termes répétées exactement sans emploi anaphorique ou de synonymes (Burstein, 2009; Gurevych *et al.*, 2009).

La lisibilité dépend aussi d'une cohérence discursive, au moins à courte portée. Pour cela il faut détecter des ruptures thématiques inappropriées entre des unités textuelles adjacentes (Burstein, 2009). Détecter ces ruptures peut être réalisé par le biais de méthodes utilisant une représentation vectorielle et lexicale du sens. Le Text Tiling de Martin Hearst (Hearst, 1997) permet à l'aide de telles représentations de réaliser un découpage des paragraphes en ensembles de phrases cohérents.

Une dissertation n'est bien rédigée que si elle met en valeur une thèse soutenue par suffisamment d'arguments, eux-mêmes étayés par des faits. Une argumentation insuffisamment alimentée et structurée peine à convaincre et ne répond pas non plus à une nécessité d'exposer des notions attendues. Pouvoir construire un discours complet en restituant des connaissances organisées est un moyen de montrer l'acquisition, la compréhension et la maîtrise de notions (Trausan-Matu, 2010b). Cela permet d'aborder un point concernant l'évaluation de la compréhension voire de vérifier un caractère de vérité des propos.

C'est pourquoi, il est possible de considérer que la rédaction d'un essai respectant les points 2 et 3 est un indicateur de compréhension. L'appariement de ces productions avec des textes, fortement similaires, faisant autorité sur la question ou jouant le rôle d'étalons évalués par un groupe de juges humains pour chaque « note » donnée est une manière de pouvoir donner un score à la compréhension. Ce qui reviendrait à un calcul de similarité avec, par exemple, un cours (Dessus, 1999), ou une classification de nouvelles copies dans des catégories correspondant à chaque échelon de notes, par à un calcul de similarité maximale avec les copies représentatives des catégories (Foltz *et al.*, 1999). Nous avons ainsi soit une répartition des copies selon leur similarité avec un « gold-standard » soit une tâche de classification supervisée.

Ces similarités sont calculées sur le contenu (le signifiant) mais doivent rendre compte du sens (le signifié). C'est pourquoi des modèles lexicaux de calcul et de représentation du sens peuvent être adaptés (pour plus de détails, voir 3.1). En outre, l'utilisation d'ontologies, de thésaurus spécifiques à un domaine (GeneOntology [1], UMLS [2]) ou de réseaux sémantiques plus généraux (WordNet [3]) peut permettre la manipulation plus directe de concepts et de sens en faisant abstraction des lexèmes qui les incarnent, permettant notamment d'unifier des synonymes.

1. http ://www.geneontology.org/
2. http ://www.nlm.nih.gov/research/umls/
3. http ://wordnet.princeton.edu/

2.3 Aide à la lecture et à l'écriture

Lire des textes en langue non maternelle ou contenant des termes spécifiques à un domaine, abordant des concepts inconnus (« loin au-delà de la zone proximale de développement») ou rédigés avec un style pompeux peut être une tâche difficile. C'est pourquoi, ces lectures peuvent nécessiter une aide extérieure sous la forme de simplifications de textes, de propositions de synonymes, de glossaires (Gaudio, 2007) ou encore de documents tiers explicitant définitions, concepts ou simplement traitant du même sujet.

Dans l'idéal, ces documents doivent être accessibles à l'apprenant tant en terme de vocabulaire que de connaissances pré-requises, tout en permettant l'acquisition de nouvelles connaissances. Les connaissances qu'ils transportent sont alors présentes dans ce que Lev Vygotsky, père du constructivisme, nomme la zone proximale de développement, « ni trop proches » du modèle de l'apprenant « ni trop éloignés» (Zampa, 2005).

La rédaction, quant à elle, peut être assistée en fournissant des correcteurs automatiques orthographiques et syntaxiques ou encore des dictionnaires de synonymes mais aussi en permettant d'identifier les notions exposées par l'apprenant et abordées dans le cours dans le cadre duquel la production écrite s'inscrit (Lemaire et Dessus, 2001).

L'identification d'une structure des thèmes et des notions couverts par l'écrit, et calquée sur la structure typographique ou logique peut mettre en évidence des problèmes de cohérence. En observant les couvertures respectives d'unités adjacentes, on peut alors identifier des ruptures thématiques (Lemaire et Dessus, 2001). En cela, ces possibilités rejoignent le *Text-Tiling* (Hearst, 1997) permettant d'identifier ces ruptures afin de délimiter des zones textuelles cohérentes.

2.4 Gestion de contenus et apprentissage collaboratif(s)

L'essor du Web a permis l'accès à des ressources en ligne comme des sites spécialisés ou des encyclopédies numériques, mais il a aussi permis de créer un savoir construit socialement dans des forums, des blogs ou encore des wikis. Dans (Gurevych *et al.*, 2009), les auteurs insistent quasi-exclusivement sur des travaux dans lesquels le TAL est utilisé pour organiser et structurer la connaissance notamment dans des wikis. Ces travaux sont dans l'ensemble très proche de l'ingénierie des connaissances et de la recherche documentaire.

Mais il s'agit aussi d'analyser des échanges dans le cadre de débats imposés, de forums de formations ou de discussions. Le projet européen LTfLL a notamment apporté une contribution non négligeable et on peut noter l'existence du module PolyCAFe (Rebedea *et al.*, 2010; Trausan-Matu, 2010a,b) qui fournit des *feedbacks* aux différents protagonistes d'une situation d'apprentissage sous forme de débats. Nous revenons en particulier sur ce module dans la section suivante.

Entre autres, des travaux sur l'analyse automatique de forums de formation à distance alimentent aussi cette catégorie. Dans (Sidir *et al.*, 2006), ce sont les forums libres c'est-à-dire sans limite de temps et sans tâche fixe qui sont ciblés. Ces travaux essayent notamment d'identifier s'il existe « des processus de co-construction de connaissances entre apprenants indépendamment des interventions des tuteurs» .

Une analyse thématique automatique avec le logiciel ThemAgora souligne une progression discursive en rapport avec celle de la formation. Une analyse linguistique et manuelle du discours

fondée sur le modèle d'exposition de Yamada dégage plusieurs mouvements dans le discours correspondant à des phases différentes de cette co-construction de connaissance.

2.5 Discussion

Ces premiers éléments d'un état de l'art nous amènent à deux conclusions. La première est que l'utilisation du TAL dans les EIAH semble se trouver principalement dans l'Apprentissage de la Langue Assisté par Ordinateur (ALAO) et que de surcroît l'apprentissage en groupe n'est peut-être pas ce qui tire le plus parti du TAL. La seconde conclusion réside dans le fait que toutes les axes linguistiques ne sont pas couverts par l'application du TAL aux EIAH. En effet, la dimension rhétorique et l'*argumentative zoning* (Teufel, 1999) semblent délaissés et l'utilisation de modèles discursifs ou dialogiques encore embryonnaires. C'est pourquoi, le TAL a sa place dans les EIAH car il pourrait notamment apporter des indicateurs qualitatifs sur la construction de connaissances, ce qui fait actuellement défaut, laissant donc un espace encore vierge entre TAL et EIAH et de nombreuses opportunités (Antoniadis, 2008; Burstein, 2009; Antoniadis *et al.*, 2009)

3 Du cognitif/connexionniste au socio-constructiviste : de LSA vers le discours et le dialogue au travers d'un réseau

3.1 L'Analyse Sémantique Latente : d'un modèle documentaire, à une représentation des connaissances et un modèle de leurs acquisitions

L'Analyse Sémantique Latente est un modèle statistique et lexical de représentation vectorielle du sens latent des termes porté par les relations de co-occurence locale qu'ils entretiennent au sein de documents d'un corpus (Deerwester *et al.*, 1990) .

Un corpus de documents est représenté par une matrice de co-occurrence qui associe à chacun de ces documents (ou tout autre unité textuelle de grain pertinent) le nombre d'occurrences de chacun des termes du corpus. Une décomposition en valeurs singulières permet de réduire le rang de la matrice aux termes les pertinents et d'obtenir des vecteurs de mêmes dimensions. Ce modèle permet ainsi de calculer la proximité sémantique entre termes ou documents, c'est-à-dire les contextes qu'ils partagent directement ou indirectement, grâce à un simple cosinus entre vecteurs.

D'abord appliqué dans le champ de la recherche documentaire (Dumais, 1991), ce modèle s'est vu employé dans différents autres domaines comme l'apprentissage et l'acquisition de connaissances, et utilisé dans des EIAH fournissant certains indicateurs.

LSA permet de simuler certains processus cognitifs et notamment l'acquisition de vocabulaire par exposition à des textes. (Landauer et Dumais, 1997) ont montré qu'en faisant traiter par LSA autant de mots/textes qu'un jeune entre 2 et 20 ans, le nombre de nouveaux mots acquis par jour par celui-ci est du même ordre que le nombre de nouvelles paires de mots proches sémantiquement construites. De plus, la simulation de réponses à des tests de synonymies du TOEFL par LSA, après exposition au contenu d'une encyclopédie montre des résultats proches de

ceux atteints par une population d'élèves étrangers (Landauer et Dumais, 1997). La simulation de réponses de LSA à des QCM sur des notions de mathématiques après traitement de quelques cours obtient des résultats moins brillants. Dans ces deux cas, les réponses choisies sont celles qui sont les plus proches sémantiquement des énoncés dans le modèle vectoriel calculé à partir du corpus d'apprentissage. La différence de performance met alors en évidence l'importance de ce corpus pour ces processus décisionnels. LSA a aussi été mis à profit dans des tâches de notations et d'évaluation de productions écrites individuelles (Foltz *et al.*, 1999). Dans (Dessus et Lemaire, 2002) les auteurs ont procédé à des expériences similaires mais avec un traitement différent. Une indexation thématique sur deux niveaux hiérarchiques (Sujets et notions) de cours de sociologie permet d'évaluer ces productions tant en termes de cohérence que de couverture du sujet.

En effet, le traitement de ce corpus par LSA selon ces deux niveaux permet de fournir des indicateurs sur la couverture du sujet selon un axe macroscopique et microscopique. Le calcul de similarité par LSA entre la production de l'apprenant et les différents sujets et notions du cours permet de déterminer quel est le contenu du cours couvert par sa copie d'une part, mais aussi d'appréhender le plan, l'organisation de sa copie grâce à un appariement à des grains plus petits. Ces indications de couvertures au niveau microscopique permettent aussi de fournir un retour sur la cohérence textuelle inter-phrastique (Foltz *et al.*, 1999; Dessus, 1999, 2000), d'une manière proche du *Text-Tiling* (Hearst, 1997), et d'identifier des ruptures et des changements thématiques brutaux et inattendus faisant baisser la qualité de la copie.

Ce principe est intégré dans le logiciel Apex (Dessus et Lemaire, 2002) qui permet de guider l'apprenant dans l'exploration de documents ou de cours à des fins d'apprentissage. Ce système propose à un étudiant de lire des textes en rapport avec une requête qu'il fournit puis de dire s'il peut ou non résumer ce texte. Dans le cas affirmatif, il est invité à construire un résumé du texte qui est alors évalué par le processus précédent. Dans l'autre cas ou si le résumé est de faible qualité, un autre texte est proposé. Le choix du texte suivant est crucial et fondé sur LSA. En effet, le texte suivant proposé à l'apprenant est celui qui est le plus proche sémantiquement de l'ensemble des résumés qu'il a pu produire.

Deux choses sont intéressantes dans cette utilisation de LSA. La première est le fait que les modèles de représentation des connaissances de l'apprenant et celui du but à atteindre sont les mêmes : LSA. L'état des connaissances de l'apprenant est alors l'ensemble des résumés qu'il a pu produire et le but réside dans les documents proposés en réponse à sa requête préalable. L'autre aspect intéressant est d'utiliser la réponse de l'apprenant sur sa capacité à résumer un texte pour en réalité savoir s'il l'a compris.

Nous avons essayé de présenter l'Analyse Sémantique Latente, certaines de ses possibilités et de ses utilisations en rapport avec le domaine éducatif. Nous voulions insister sur le fait qu'il est ici principalement utilisé pour représenter l'état de connaissance d'un individu et permet entre autres de comparer ses productions à un matériau qui fait autorité sur cette connaissance, grâce une représentation des relations entre termes. À notre connaissance, LSA ne semble pas employée dans des situations d'apprentissage qui sont le siège d'interactions distantes et informatiquement médiatisées entre apprenants, du moins pas à même escient.

3.2 Vers une construction sociale de connaissances grâce au dialogue et au partage dans un réseau

Nous voudrions présenter des travaux mettant en œuvre différentes techniques du TAL pour restituer certains indicateurs dans la collaboration. Leurs travaux s'appuient sur les transpositions des notions de dialogisme, de polyphonie et d'inter-animation de Bakhtin ainsi que sur l'hypothèse que discours et dialogue jouent un rôle prépondérant dans la construction et l'acquisition des connaisances. (Trausan-Matu, 2010b).

3.2.1 Définitions

Voici les définitions des notions relatives à Bakhtin traduites depuis (Trausan-Matu, 2010b) :

Dialogisme Un concept introduit par Mikhail Bahthin, qui considère que chaque création et activité langagière humaine est un dialogue, incluant non seulement les conversations mais aussi des textes ou même des réflexions.

Inter-animation Un phénomène spécifique à la polyphonie ou à des groupes de personnes collaborant dans lequel plusieurs voix entrent en dialogue et, à cause d'interactions caractérisées par le même ou le différent (centripète ou centrifuge), un thème est développé.

Polyphonie Une réalisation conjointe qui implique plusieurs individus qui construisent en collaborant une structure cohérente et durable à partir d'un thème donné, même si des dissonances délibérées et transitoires peuvent apparaitre. Afin d'atteindre une cohérence, différentes règles assurant l'harmonie se doivent d'être respectées.

3.2.2 Le système

Dans (Trausan-Matu, 2010a) et (Rebedea *et al.*, 2010), les auteurs présentent un système nommé PolyCAFe analysant les échanges entre des étudiants dans une optique de débat, et de synthèse, concernant un domaine bien défini, sur une plateforme informatique dédiée conservant les traces de ces discussions.

Ce système associe à une chaîne de traitement linguistique traditionnelle une ontologie représentant les concepts du domaine, ici les interfaces Homme-Machine. Afin d'éviter des ambiguïtés dues à la polysémie des langues naturelles et dans le but d'identifier les différents fils de discussion, une LSA est réalisée au préalable sur un corpus du domaine comparant les concepts évoqués dans deux messages au sein de l'espace sémantique construit.

Le but est ici d'identifier les dimensions longitudinale et transversale de la polyphonie mais aussi de rendre compte de l'inter-animation : l'entremêlement des propos des uns dans ceux des autres et d'identifier les références des uns aux autres. À cette fin, le système de discussion invite les participants à préciser à quel apprenant ils répondent. Cette information est utilisée pour identifier de premières interactions explicites. Des traitements linguistiques de plus haut niveau prennent place pour identifier les références implicites des uns aux autres. Ces traitements consistent notamment en une identification de répétitions, une résolution de la coréférence, un calcul de similarité grâce à LSA et la prise en compte de connecteurs logiques afin d'identifier des actes de langage et des paires adjacentes. Cette dernière identification peut permettre de détecter des comportements et les différents rôles qu'endossent les apprenants dans la discussion.

Ce point avait déjà été envisagé dans (George, 2004) mais évincé suite à des réserves concernant la faisabilité d'une automatisation de cette détection.

Capter les différentes interactions entre les apprenants permet de construire alors un réseau qui va représenter l'inter-animation amenant à une construction du savoir. Une analyse des réseaux sociaux identifie différents critères significatifs comme la centralité des graphes, les degrés, les participants faisant autorité (au sens du pagerank de Google) ou la cohésion, notamment avec le calcul de composantes fortement connexes et de cliques. Cette étape permet de retourner des indicateurs quant à la participation de chacun dans les différents fils de discussions ou la position plus ou moins centrale dans les débats. S'ajoutent à ces indicateurs des informations sur la lisibilité et la cohérence textuelle des propos.

Ces travaux sont très intéressants car ils mettent en perspective les techniques actuelles du TAL de bas niveau (analyse morphosyntaxique, LSA, ontologie) et de haut niveau (discours et dialogisme) mais aussi les techniques d'analyse de réseaux sociaux pour des situations collaboratif en ligne.

3.3 Bilan

Nous avons voulu montrer qu'en passant de l'individu au groupe, certaines méthodes restaient intéressantes même si leur utilisation n'était pas identique et que les problématiques concernant la construction de la connaissance se déplacent. Chez l'individu, c'est la connaissance construite qui nous intéresse alors que dans une co-construction, ce sont aussi les constructeurs et le chantier. De plus, une constante apparait : la construction du sens grâce, non pas aux éléments (termes ou individus) eux-mêmes, mais grâce aux relations qu'ils entretiennent entre eux.

4 Réel et Focalisation

Nous cherchons à explorer un champ de recherche pluridisciplinaire mariant EIAH et TAL. Cela nous impose de résoudre un conflit méthodologique. En effet, le champ des EIAH aborde l'ingénierie pour répondre à des besoins et permettre des usages par un média et un outillage informatique à des fins didactiques : il met en œuvre des analyses théoriques et généralistes, sans nécessairement de matériau d'observation, des modélisations pédagogiques et informatiques, des expérimentations et des évaluations par les utilisateurs des systèmes produits. Le TAL utilise des méthodologies plus empiriques mettant habituellement à profit l'observation d'un matériau représentatif d'une entrée à traiter, proposant des modèles informatiques et linguistiques éprouvés dans le cadre d'expérimentations sur des corpus d'évaluations. La vastitude apparente de notre problématique nous invite ainsi à amorcer nos recherches en précisant la tâche à traiter c'est-à-dire la situation ou les situations d'apprentissages, les protagonistes, les traces textuelles collectées et les indicateurs à produire, mais aussi à observer un matériau langagier réel à partir d'un corpus de traces.

Les situations d'apprentissage socio-constructivistes auxquelles nous nous intéressons et pour lesquelles nous devons fournir des indicateurs restent floues. Nos discussions ont mis en avant parmi ces situations celles d'apprentissages par projet ou de débat/résolutions de problèmes. Dans ces deux situations, les apprenants sont amenés à discuter, à faire des recherches personnelles, à partager les informations recueillies et leurs points de vue. Il s'agit aussi de critiquer le point

de vue d'autrui ou le sien, de négocier le sens de certaines informations, la validité de propos, d'exclure en accord des informations ou encore de discuter en parallèle de plusieurs notions. Ce type d'exercice attend généralement sa clôture par une étape de synthèse et de conclusion des débats. Cette étape est le siège de consensus ou de l'intégration de plusieurs opinions et a pour vocation de répondre à l'exercice. Cette réponse peut prendre plusieurs formes : clore les échanges ou être rédigée collectivement de manière déportée. Dans ce dernier cas, l'élaboration de la réponse se déroule soit en parallèle des échanges soit *a posteriori* et peut alors être révisée maintes et maintes fois. Des aspects comme la discipline abordée, la durée des exercices et le nombre d'apprenants participant au même exercice restent obscurs. Il nous faudra aussi décider si nous devons traiter oui ou non ces paramètres de manière générique, ce qui paraît ambitieux.

Nous n'avons pas non plus défini ou décrit la population qui constitue les différents utilisateurs, la discipline concernée, leur niveau dans leur discipline la langue dans laquelle ils discutent (maternelle ou seconde), leur niveau de langue. Or ces informations sont nécessaires pour établir un profil, un modèle de l'apprenant et pourraient s'avérer déterminantes dans une chaine de traitement automatique de la langue.

Les traces textuelles que nous allons traiter sont à la fois des productions qui répondent aux exercices auxquels prennent part les apprenants, et des messages échangés au travers d'un réseau de manière synchrone ou asynchrone. Ce type de messages présente des caractéristiques proches de l'oral et peut notamment être bruité. (Sidir *et al.*, 2006; Bouchet et Sansonnet, 2006).

Nous travaillons donc un matériau réel et bruité dans lesquels les usages et les normes ont plus leur place que des règles. Pour appréhender cette « réalité empirique » et capter certains invariants propres à ce genre et à ce type de discours, il est donc nécessaire de s'imprégner d'un échantillon représentatif et de s'atteler à une description. Notre tâche consiste à manipuler du sens, or, selon Rastier (Rastier, 2005), le sens obéit notamment à ces spécificités à causes des problèmes sémantiques que sont l'implicite et la polysémie, caractérisant la langue naturelle.

Ainsi, (Rastier, 2005; Williams, 2003; Bouchet et Sansonnet, 2006) nous incitent donc à construire un corpus d'étude pouvant nous aider à focaliser nos recherches sur le réel d'une situation. Même si ce n'est qu'une amorce, cela nous permettra peut-être par la suite d'atteindre une certaine généricité attendue.

Conclusion

Nous avons présenté la problématique que nous formulons, qui consiste à explorer les capacités du TAL pour l'extraction d'indicateurs sur la construction collective de connaissances au sein d'EIAH, et avons tenté de la justifier en présentant des opportunités et en montrant la couverture partielle des EIAH par le TAL et du TAL par les EIAH. Nous avons montré les possibilités qu'offre LSA pour représenter et évaluer les connaissances d'un individu, mais aussi un système récent mettant en œuvre différentes techniques du TAL pour fournir des indicateurs quantitatifs sur la construction collective de connaissances et qui représentent un point de départ et de repère pour nos travaux. Nous terminons en présentant les particularités liées à la pluridisciplinarité de notre problématique et des choix qu'il nous faudra peut-être faire assez tôt malgré un conflit méthodologique qui s'y oppose.

Nos perspectives concernent la consolidation de l'état de l'art et la justification de notre problé-

matique mais aussi des questions liées à la représentation de la construction et son évolution en prenant en compte ce qui est construit, qui construit quoi, et si le multiple construit du même ou du différent. Il nous faut aussi approfondir les moyens d'identifier des moments du discours et de calculer des différentiels entre l'état de la construction des connaissances entre deux instants ou en un instant et état cible. Les graphes (Zouaq *et al.*, 2000) et les cartes conceptuelles nous semblent un moyen envisageable (Berlanga *et al.*, 2009).

Références

ANTONIADIS, G. (2008). *Du TAL et de son apport aux systèmes d'apprentissage des langues : Contributions*. Habilitation à diriger des recherches en informatique et traitement automatique des langues, Université Stendhal - Grenoble 3.

ANTONIADIS, G., GRANGER, S., KRAIF, O., PONTON, C., MEDORI, J. et ZAMPA, V. (2009). Integrated Digital Language Learning. *In* BALACHEFF, N., LUDVIGSEN, S., JONG, T., LAZONDER, A. et BARNES, S., éditeurs : *Technology-Enhanced Learning*, chapitre 6, pages 89–103. Springer Netherlands, Dordrecht.

BERLANGA, A. J., KALZ, M., STOYANOV, S., van ROSMALEN, P., SMITHIES, A. et BRAIDMAN, I. (2009). Using language technologies to diagnose learner's conceptual development. *Advanced Learning Technologies, IEEE International Conference on*, 0:669–673.

BOUCHET, F. et SANSONNET, J. (2006). Étude d'un corpus de requêtes en langue naturelle pour des agents assistants. *In Actes du Deuxième Workshop sur les Agents Conversationnels Animés (WACA 2006)*, pages 95–104, Toulouse, France.

BURSTEIN, J. (2009). Opportunities for natural language processing research in education. *CICLing 09 Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, 5449:6–27.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. et HARSHMAN, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

DESSUS, P. (1999). Apex, un système d'aide à la préparation d'examens. *Sciences et Techniques éducatives*, 6(2):409–415.

DESSUS, P. (2000). Construction de connaissances par exposition à un cours avec LSA. *In Cognito*, 18:27–34.

DESSUS, P. et LEMAIRE, B. (2002). *Using production to assess learning : An ILE that fosters self-regulated learning*, volume 2363, pages 772–781. Springer.

DUMAIS, S. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods*, 23(2):229–236. 10.3758/BF03203370.

FOLTZ, P. W., LAHAM, D. et LANDAUER, T. K. (1999). Automated essay scoring : Applications to education technology. *In Proceedings of EDMEDIA*, volume 1, pages 939–944. AACE.

GAUDIO, R. D. (2007). Supporting e-learning with automatic glossary extraction : Experiments with portuguese. *In RANLP 2007 workshop : Natural Language Processing and Knowledge Representation for eLearning Environments*.

GEORGE, S. (2004). Analyse automatique de conversations textuelles synchrones d'apprenants pour la détermination de comportements sociaux. *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation (STICEF)*, Vol. 10:p. 165–193.

GUREVYCH, I., BERNHARD, D. et BURCHARDT, A. (2009). Educational natural language processing. Notes for ENLP tutorial held at AIED 2009 in Brighton.

HEARST, M. A. (1997). Texttiling : Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, pages 33–64.

HOU, H.-T. et WU, S.-Y. (2011). Analyzing the social knowledge construction behavioral patterns of an online synchronous collaborative discussion instructional activity using an instant messaging tool : A case study. *Computers & Education*, 57:1459–1468.

KARAMANIS, N. (2006). Generating multiple-choice test items from medical text : A pilot study. *In In Proceedings of INLG 2006*, pages 104–107.

LANDAUER, T. K. et DUMAIS, S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.

LEE, J. et SENEFF, S. (2007). Automatic generation of cloze items for prepositions. *In INTERSPEECH*, pages 2173–2176. ISCA.

LEMAIRE, B. et DESSUS, P. (2001). A system to assess the semantic content of student essays. *Journal of Educational Computing Research*, 24(3):305–320.

RASTIER, F. (2005). *Enjeux épistémologiques de la linguistique de corpus*, pages 31–45. Presses Universitaires de Rennes.

REBEDEA, T., DASCALU, M., TRAUSAN-MATU, S., BANICA, D., GARTNER, A., CHIRU, C. et MIHAILA, D. (2010). Overview and preliminary results of using polycafe for collaboration analysis and feedback generation. *In Proceedings of the 5th European conference on Technology enhanced learning conference on Sustaining TEL : from innovation to learning and practice*, EC-TEL'10, pages 420–425, Berlin, Heidelberg. Springer-Verlag.

SIDIR, M., LUCAS, N. et GIGUET, E. (2006). De l'analyse des discours à l'analyse structurale des réseaux sociaux : une étude diachronique d'un forum éducatif. *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation (STICEF)*, 13.

TEUFEL, S. (1999). *Argumentative Zoning : Information Extraction from Scientific Text*. Thèse de doctorat, University of Edinburgh, School of Cognitive Science.

TRAUSAN-MATU, S. (2010a). Automatic support for the analysis of online collaborative learning chat conversations. *In Proceedings of the Third international conference on Hybrid learning*, ICHL'10, pages 383–394, Berlin, Heidelberg. Springer-Verlag.

TRAUSAN-MATU, S. (2010b). The polyphonic model of hybrid and collaborative learning. *In* WANG, F. L., FONG, J. et KWAN, R., éditeurs : *Handbook of Research on Hybrid Learning Models : Advanced Tools, Technologies, and Applications*, pages 466–486. Information Science Publishing, New York.

WILLIAMS, G. (2003). Texte et Corpus. *In Actes des Troisièmes Journées de la Linguistique de Corpus*, pages 1–307.

ZAMPA, V. (2005). Utilisation de l'analyse sémantique latente pour tenter d'optimiser l'acquisition par exposition à une langue étrangère de spécialité. Volume 8.

ZOUAQ, A., FRASSON, C. et ROUANE, K. (2000). The explanation agent. *In* GAUTHIER, G., FRASSON, C. et VANLEHN, K., éditeurs : *Intelligent Tutoring Systems, 5th International Conference, ITS 2000, Montréal, Canada, June 19-23, 2000, Proceedings*, volume 1839 de *Lecture Notes in Computer Science*, pages 554–563. Springer.