L'analyse de la complexité du discours et du texte pour apprendre et collaborer

Mihai Dascalu, Philippe Dessus, Stefan Trausan-Matu

L'analyse de la complexité du discours et du texte pour apprendre et collaborer

par Mihai Dascălu, Philippe Dessus, Ştefan Trăuşan-Matu
Document de travail, le 22/09/14

Introduction

L'apprentissage collaboratif assisté par ordinateur et les technologies d'e-learning devenant de plus en plus populaires et intégrés dans des contextes éducatifs, le besoin se fait sentir de disposer *d'outils d'évaluation automatique et d'aide aux enseignants ou tuteurs* pour les deux activités, fortement couplées, de compréhension de textes et collaboration entre pairs. Bien qu'une analyse de surface de ces activités est aisément réalisable, une compréhension plus profonde et complète du discours en jeu est nécessaire, complétée par une analyse de l'information méta-cognitive disponible par diverses sources, comme par exemples les auto-explications des apprenants.

Dans ce contexte, nous utilisons un modèle dialogique issu des travaux de Bakhtine (1981, 1984) pour analyser les conversations collaboratives, et une approche théorique visant à unifier les activités de compréhension et de collaboration, en utilisant des graphes de cohésion. Plus spécifiquement, nous nous sommes centrés sur la *dimension individuelle de l'apprentissage*, analysée à partir de l'identification de stratégies de lecture et sur la mise au jour d'un modèle de la complexité textuelle intégrant des facteurs de surface, lexicaux, morphologiques, syntaxiques et sémantiques. En complément, *la dimension collaborative de l'apprentissage* est centrée sur l'évaluation de l'implication des participants, ainsi que sur l'évaluation de leur collaboration par deux modèles computationnels : un *modèle polyphonique*, défini comme l'inter-animation de voix selon de multiples perspectives, un *modèle* spécifique de *construction sociale de connaissances*, fondé sur un graphe de cohésion et un mécanisme d'évaluation des tours de parole.

Notre approche met en œuvre des techniques avancées de traitement automatique de la langue et a pour but de formaliser une évaluation qualitative du processus d'apprentissage. Ainsi, deux perspectives fortement liées sont prises en considération : d'une part, *la compréhension*, centrée sur la construction de connaissances et les auto-explications à partir desquelles les stratégies de lecture sont identifiées ; d'autre part la *collaboration*, qui peut être définie comme l'implication sociale, la génération d'idées ou de voix en interanimation dans un contexte donné.

Vue intégrée

En bref, notre objectif est de soutenir les processus de compréhension de l'apprentissage individuel et collaboratif ou, plus précisément, de soutenir les processus de construction de la connaissance personnelle et sociale sous-jacents, à travers l'utilisation des outils automatiques, notamment l'évaluation de la cohésion textuelle textes lus et des productions des apprenants. C'est selon trois perspectives que nous cherchons à remplir cet objectif (Figure 1) :

— le cycle interne modèle le processus d'apprentissage du point de vue de renforcement des connaissances (Bereiter, 2002 ; Scardamalia, 2002 ; Stahl, 2006),

— le processus d'évaluation en termes de compréhension sur la base de l'évaluation des productions des apprenants.

— des outils de Traitement Automatique des Langues (TAL), nécessaire pour effectuer l'analyse du discours (Jurafsky & Martin, 2009).

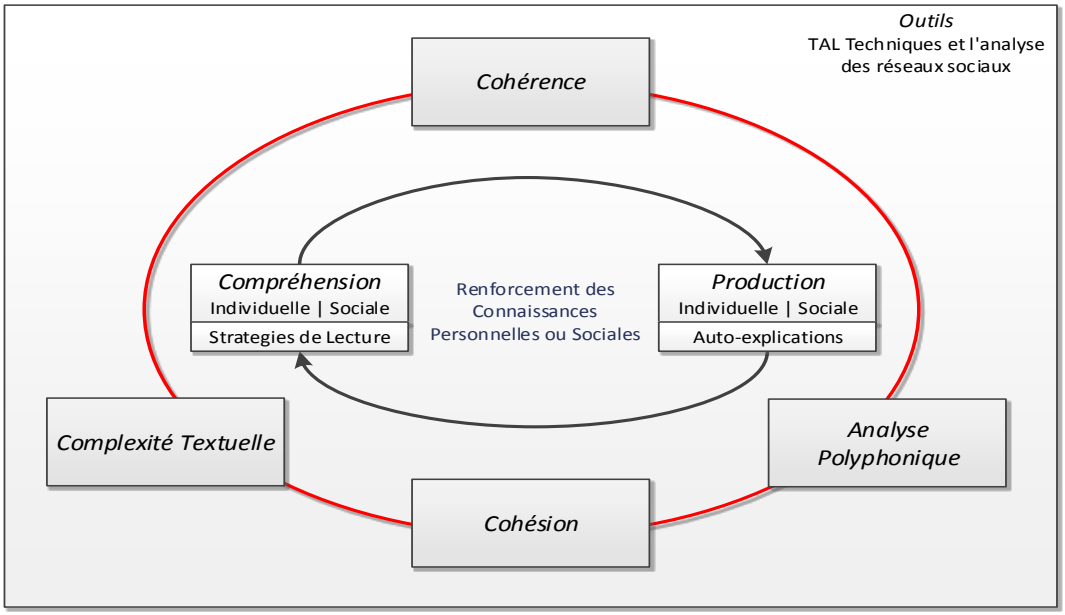

Figure 1 Vue intégrée des aspects et concepts théoriques

Les outils de traitement automatique de la langue sont l'analyse de la sémantique latente (LSA) (Deerwester et al., 1989 ; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990 ; Dumais, 2004 ; Landauer & Dumais, 1997) et l'allocation de Dirichlet latente (LDA) (Blei, Ng, & Jordan, 2003).

LSA représente sous la forme d'un espace vectoriel les relations entre termes et les documents (paragraphes) qui les contiennent, en se fondant sur l'analyse de leurs co-occurrences. Cela permet d'évaluer la similarité sémantique inter-termes et entre termes et documents (Landauer, Foltz, & Laham, 1998 ; Manning & Schütze, 1999). LDA permet de réaliser une identification de thèmes via un mécanisme d'inférence probabiliste de structures thématiques dans les documents.

Ces deux méthodes sont de type « paquets-de-mots » (bag of words), et ne tiennent pas compte de l'ordre des mots dans les phrases et documents. Cela est bien entendu une grande approximation, mais qui n'est pas trop gênante puisqu'il s'agit de récupérer les mots-clés principaux et des indices de similarités entre ces mots-clés, apparaissant dans un grand corpus.

ReaderBench, présentation générale

ReaderBench (Dascalu, 2014 ; Dascalu, Dessus, Bianco, Trausan-Matu, & Nardy, 2014) permet d'appliquer ces techniques sur des textes narratifs (histoires) ou sur des conversations collaboratives, en particulier dans des chats ou des forums de discussion (Dascalu, Trausan-Matu, & Dessus, 2014 ; Nistor et al., 2014) et donc dans des scénarios pédagogiques plus complexes. *ReaderBench* peut être diffusé et utilisé à des fins de recherche. Le développement de ce logiciel a été partiellement réalisé par l'ANR (projet DEVCOMP) et les projets POSDRU/107/1.5/S/76909, POSDRU/159/1.5/S/134398, 264207 ERRIC FP7-REGPOT-2010-1 et FP7 2008-212578 LTfLL.

L'utilisation de stratégies pendant la lecture est un facteur important de la compréhension chez les adultes comme chez les enfants et peuvent être recueillies par les explications à haute voix des élèves, au fur et à mesure qu'ils lisent un texte. Des recherches les ont catégorisées de cette manière : régulation de la compréhension (« *Je crois que j'ai compris que le héros hésite à blesser le monstre.* »), paraphrase (répétition presque à l'identique d'une proposition), élaboration (construction d'une

nouvelle connaissance à partir du texte), prédiction (« *Le héros va se marier avec la princesse* ») et *la mise en relation* (« *Mais le héros s'est déjà battu avec le monstre au tout début* »).

L'analyse automatique de l'emploi de ces stratégies à partir des explications des élèves (transcrites de l'oral) est présenté sur la Figure 2 (copie d'écran de *ReaderBench*). On y voit des mesures de comparaison sémantique entre les paragraphes du texte à lire (dernière colonne de droite), et des catégories de lecture pour chaque explication (en gris), codées ainsi : régulation, paraphrase [entre crochets, un numéro d'index référant aux mots du texte lu], élaboration [*] et *mise en relation*. Ainsi, l'enseignant peut avoir une idée précise du niveau de compréhension du texte lu par *Mathilda*. On peut par exemple noter qu'un nombre trop important de paraphrases (répétition de portions de texte lu sans nécessaire compréhension profonde), au détriment d'élaborations ou mises en relation, indique une compréhension de plus haut niveau.

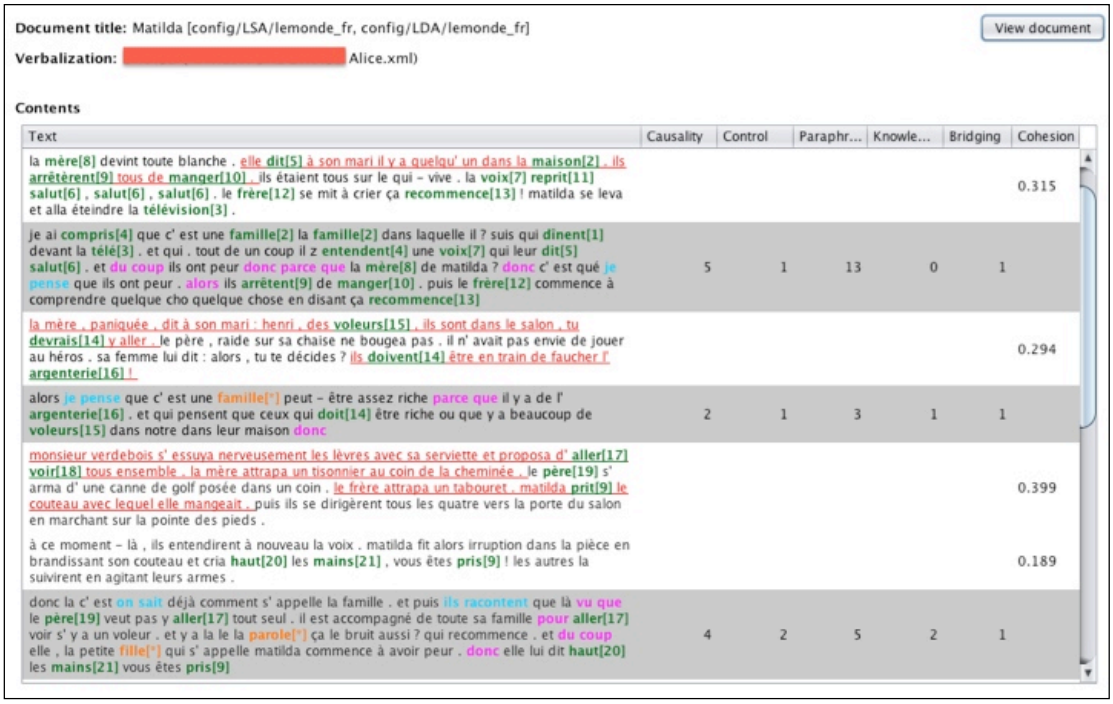

Figure 2 Analyse des stratégies de lecture et de la cohésion textuelle dans *ReaderBench*

Scénarios

La Figure 3 présente les activités de chacun et leur succession du point de vue de *l'apprentissage individuel*. L'enseignant commence par utiliser le logiciel pour sélectionner des documents à lire compatibles avec le niveau de ses élèves. Ces derniers les lisent, et réalisent certaines productions suite à leur lecture, productions qui seront ensuite évaluées par l'enseignant. Ensuite, quatre boucles sont représentées. La *boucle de lecture* permet à l'apprenant de prendre connaissance du matériel de cours ; *la boucle d'écriture* permet à l'apprenant d'auto-expliquer ce qu'il a compris du cours ; et la *boucle de sélection de thème* permet à l'apprenant de saisir quelques mots-clés et de sélectionner les phrases importantes du document lu. Une boucle concerne l'enseignant et l'évaluation de la production de l'apprenant, aidé en cela par le système qui produit différentes évaluations consultables par l'enseignant.

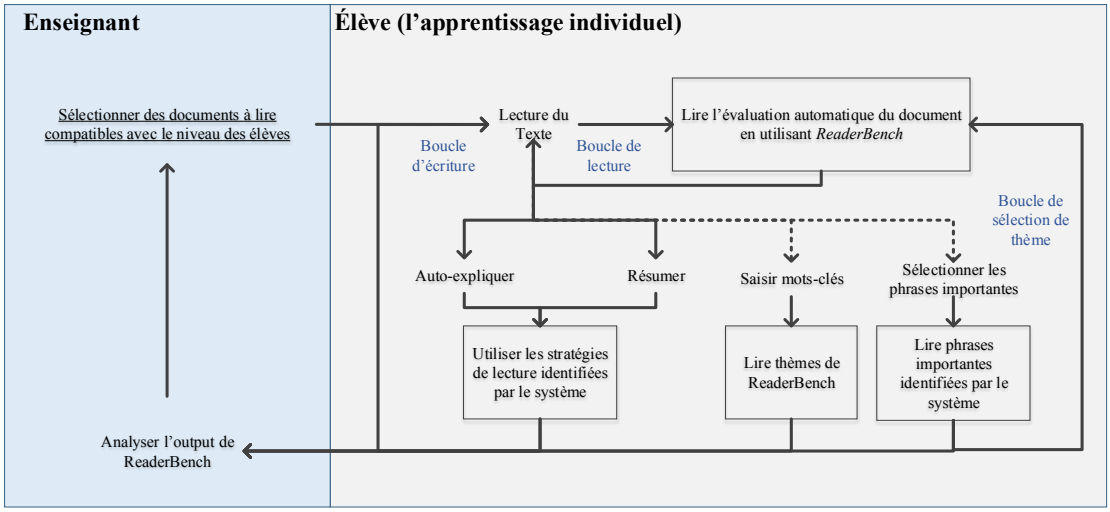

Figure 3. Représentation graphique du scénario utilisant *ReaderBench, centré sur l'apprentissage individuel, selon la perspective de l'apprenant et de l'enseignant*.

La **Figure 4** représente les boucles d'activité dans lesquelles les apprenants et l'enseignant sont engagés dans un scénario collaboratif. Deux types de boucles sont possibles. Une *boucle de lecture,* dans laquelle les apprenants prennent connaissance des interventions de leurs pairs, ainsi que d'une vue d'ensemble de la conversation (fils de discussion de forums ou clavardage). Une *boucle d'écriture*, qui met automatiquement en évidence le niveau de participation et collaboration d'un apprenant donné au sein de la conversation. De plus, l'enseignant peut pré-sélectionner et attribuer un matériel d'apprentissage à des apprenants compatibles avec leur niveau de lecture (analyse de la complexité textuelle). Bien évidemment, des phases individuelles peuvent alterner avec des phases collaboratives de manière à créer des scénarios plus complexes, utilisables dans des classes ou dans des contextes d'e-learning.

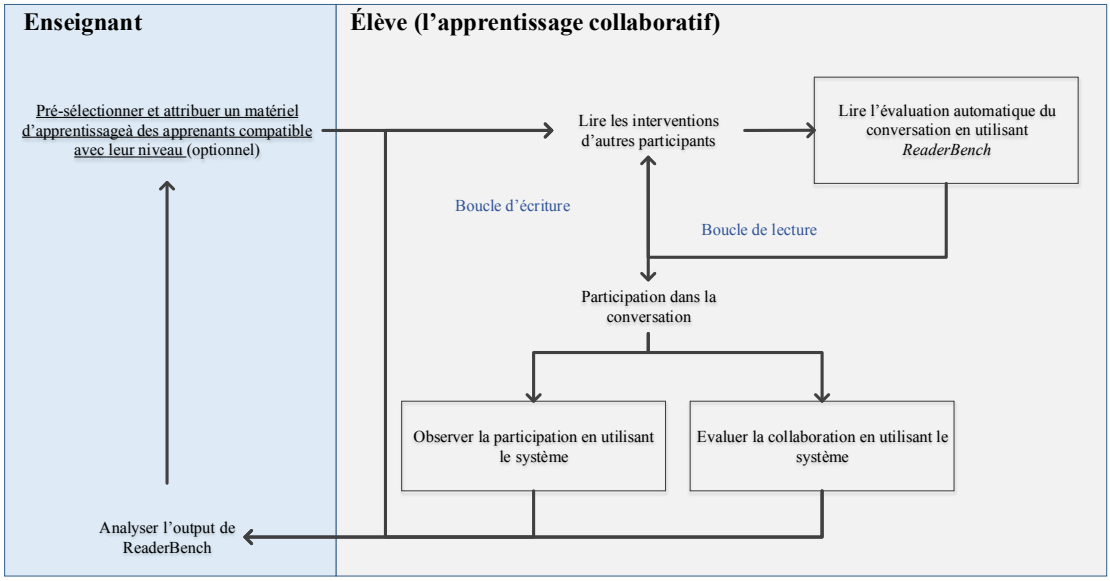

Figure 4. Représentation graphique du scénario utilisant *ReaderBench, centré sur l'apprentissage collaboratif, selon la perspective de l'apprenant et de l'enseignant.*

Conclusions

L'un des buts principaux de notre modèle est de favoriser la compréhension vue en tant que « médiatrice de l'apprentissage », en procurant des rétroactions automatiques aux apprenants et

enseignants ou tuteurs. Leur avantage est triple : leur flexibilité, leur extensibilité et, cependant, leur spécificité, car ils couvrent de multiples étapes de l'activité d'apprentissage, de la lecture de matériel d'apprentissage à l'écriture de synthèses de cours en passant par la discussion collaborative de contenus de cours et la verbalisation métacognitive de jugements de compréhension, afin d'obtenir une perspective complète du niveau de compréhension et de générer des rétroactions appropriées sur le processus d'apprentissage collaboratif.

Finalement, notre intention est d'obtenir deux axes qui se croisent, chacun dominés par un de nos axes de recherche interdisciplinaire : un axe orienté sur la *psychologie cognitive* de la cohésion et de la cohérence, alors que l'autre est plus orienté sur *l'informatique*, en mettant l'accent sur l'analyse du discours et la complexité textuelle. De plus, la compréhension utilisée pour soutenir l'apprentissage individuel ou collaboratif et les productions de l'apprenant est au centre de notre analyse, car il peut être représenté sous les productions des apprenants comme des superpositions de cohésion, cohérence, complexité textuelle et polyphonie.

Remerciements

Nous tenions à particulièrement remercier Maryse Bianco et Aurélie Nardy pour leurs conseils et soutien. Certaines parties de ce papier proviennent de l'article publié sur le blog Pole Grenoble Cognition – http://www.grenoblecognition.fr/index.php/actualites2/9-communiques/167-readerbench-un-outil-pour-evaluer-la-complexite-de-textes-et-identifier-les-strategies-de-lecture.

Références

Bakhtin, M.M. (1981). *The dialogic imagination : Four essays* (C. Emerson & M. Holquist, Trans.). Austin and London : The University of Texas Press.

Bakhtin, M.M. (1984). *Problems of Dostoevsky's poetics* (C. Emerson, Trans. C. Emerson Ed.). Minneapolis : University of Minnesota Press.

Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ : Lawrence Erlbaum Associates.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993–1022.

Chi, M.T.H., de Leeuw, N., Chui, M.H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.

D'Andrea, A., Ferri, F., & Grifoni, P. (2009). An Overview of Methods for Virtual Social Network Analysis. In A. Abraham, A. E. Hassanien & V. Snáše (Eds.), *Computational Social Network Analysis : Trends, Tools and Research Advances* (pp. 3–26). London, UK : Springer.

Dascalu, M. (2014). *Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence* (Vol. 534). Switzerland : Springer.

Dascalu, M., Dessus, P., Bianco, M., & Trausan-Matu, S. (2014). *Are Automatically Identified Reading Strategies Reliable Predictors of Comprehension ?* Paper presented at the 12th Int. Conf. on Intelligent Tutoring Systems (ITS 2014), Honolulu, USA.

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining texts, learners productions and strategies with ReaderBench. In A. Peña-Ayala (Ed.), *Educational Data Mining : Applications and Trends* (pp. 335–377). Switzerland : Springer.

Dascalu, M., Trausan-Matu, S., & Dessus, P. (2014). *Validating the Automated Assessment of Participation and of Collaboration in Chat Conversations.* Paper presented at the 12th Int. Conf. on Intelligent Tutoring Systems (ITS 2014), Honolulu, USA.

Deerwester, S., Dumais, S.T., Furnas, G.W., Harshman, R., Landauer, T.K., Lochbaum, K., & Streeter, L. (1989). USA Patent No. 4,839,853. 4,839,853 : USPTO.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Dumais, S.T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1), 188–230.

François, T., & Miltsakaki, E. (2012). *Do NLP and machine learning improve traditional readability formulas ?* Paper presented at the First Workshop on Predicting and improving text readability for target reader populations (PITR2012), Montreal, Canada.

Graesser, A.C., McNamara, D.S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iStart. *Educational Psychologist, 40*(4), 225–234.

Jurafsky, D., & Martin, J.H. (2009). *An introduction to Natural Language Processing. Computational linguistics, and speech recognition* (2nd ed.). London : Pearson Prentice Hall.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Landauer, T.K., Foltz, P.W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*(2/3), 259–284.

Linell, P. (2009). *Rethinking language, mind, and world dialogically : Interactional and contextual theories of human sense-making*. Information Age Publishing : Charlotte, NC.

Manning, C.D., & Schütze, H. (1999). *Foundations of statistical Natural Language Processing*. Cambridge, MA : MIT Press.

McNamara, D.S. (2004). SERT : Self-Explanation Reading Training. *Discourse Processes, 38*, 1–30.

McNamara, D.S., Boonthum, C., & Levinstein, I.B. (2007). Evaluating self-explanations in iSTART : Comparing word-based and LSA algorithms. In T. K. Landauer, D. S. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 227–241). Mahwah, NJ : Erlbaum.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. (2010). Coh-Metrix : Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292–330.

Millis, K., & Magliano, J.P. (2012). Assessing comprehension processes during reading. In J. P. Sabatini, E. R. Albro & T. O'Reilly (Eds.), *Assessing reading in the 21st century : Aligning and applying advances in the reading and measurement sciences* (pp. 35–53). Lanham, MD : Rowman & Littlefield Publishing.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). Measures of text difficulty : Testing their predictive value for grade levels and student performance. Washington, DC : Council of Chief State School Officers.

Nistor, N., Baltes, B., Smeaton, G., Dascalu, M., Mihaila, D., & Trausan-Matu, S. (2014). Participation in virtual academic communities of practice under the influence of technology acceptance and community factors. A learning analytics application. *Computers in Human Behavior, 34*, 339–344. doi : 10.1016/j.chb.2013.10.051

Nistor, N., & Fischer, F. (2012). Communities of practice in academia : Testing a quantitative model. *Learning, Culture and Social Interaction, 1*(2), 114–126.

Rebedea, T., Dascalu, M., Trausan-Matu, S., Banica, D., Gartner, A., Chiru, C.G., & Mihaila, D. (2010). *Overview and preliminary results of using PolyCAFe for collaboration analysis and feedback generation.* Paper presented at the Sustaining TEL : From Innovation to Learning and Practice – 5th European Conference on Technology Enhanced Learning (EC-TEL 2010), Barcelona, Spain.

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith & C. Bereiter (Eds.), *Liberal Education in a Knowledge Society* (pp. 67–98). Chicago : Open Court Publishing.

Stahl, G. (2006). *Group cognition. Computer support for building collaborative knowledge*. Cambridge, MA : MIT Press.

Trausan-Matu, S., Dascalu, M., & Dessus, P. (2012). *Textual complexity and discourse structure in Computer-Supported Collaborative Learning.* Paper presented at the 11th Int. Conf. on Intelligent Tutoring Systems (ITS 2012), Chania, Grece.

Trausan-Matu, S., Dascalu, M., & Rebedea, T. (2012). *A system for the automatic analysis of Computer-Supported Collaborative Learning chats.* Paper presented at the 12th IEEE Int. Conf. on Advanced Learning Technologies (ICALT 2012), Rome, Italy.

Trausan-Matu, S., Dascalu, M., & Rebedea, T. (2014). PolyCAFe – Automatic support for the analysis of CSCL chats. *International Journal of Computer-Supported Collaborative Learning, 9*(2), 127–156. doi : 10.1007/s11412-014-9190-y

Trausan-Matu, S., Stahl, G., & Sarmiento, J. (2006). *Polyphonic Support for Collaborative Learning.* Paper presented at the Groupware : Design, Implementation, and Use, 12th International Workshop (CRIWG 2006), Medina del Campo, Spain.

van Dijk, T.A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY : Academic Press.