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

Mihai Dascalu

UNIVERSITÉ DE GRENOBLE

THÈSE en COTUTELLE entre

- l'Université Grenoble Alpes, France

Et

- l'Université « Politehnica » de Bucarest, Roumanie

Pour obtenir le grade de

DOCTEUR

Spécialité: **Sciences de l'Education** et **Informatique et Technologie de l'Information**

Arrêté ministériel: 7 août 2006

Présentée par

Mihai DASCĂLU

Thèse dirigée par Philippe DESSUS et par Ștefan TRĂUȘAN-MATU

préparée à l'Université de Grenoble au sein du **Laboratoire des Sciences de l'Education** dans **l'École Doctorale Sciences de l'Homme, du Politique et du Territoire (ED SHPT n°454)**

préparée à l'Université « Politehnica » de Bucarest au sein du **Laboratoire de Systèmes Collaboratifs de Renforcement de Connaissances (K-Teams)** dans **l'École Doctorale Automatique et Informatique**

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

Analyzing Discourse and Text Complexity for Learning and Collaborating

Thèse soutenue publiquement **le 4 Juin 2013** devant le jury composé de:

M. Ștefan TRĂUȘAN-MATU
Professeur en Informatique, Faculté d'Automatisation et Ordinateurs, Université Politehnica de Bucarest, Roumanie, Co-directeur de thèse

M. Philippe DESSUS
Professeur en Sciences de l'Education, Laboratoire des Sciences de l'Education, Université Grenoble Alpes, Co-directeur de thèse

M. Stefano A. CERRI
Professeur en Informatique, Laboratoire d'Informatique, de Robotique et de Micro-electronique de Montpellier, Université de Montpellier, France, Rapporteur

Mme. Adina Magda FLOREA
Professeur en Informatique, Faculté d'Automatisation et Ordinateurs, Université Politehnica de Bucarest, Président du Jury

M. Bruno De LIÈVRE
Professeur au Département des Sciences et de la Technologie Educative, Université de Mons, Belgique, Rapporteur

M. Costin PRIBEANU
Chercheur senior I, Institut National de Recherche et de Développement en Informatique – ICI Bucarest, Roumanie, Rapporteur

Abstract

With the advent and increasing popularity of Computer Supported Collaborative Learning (CSCL) and e-learning technologies, the need of *automatic assessment and of teacher/tutor support* for the two tightly intertwined activities of comprehension of reading materials and of collaboration among peers has grown significantly. Whereas a shallow or surface analysis is easily achievable, a deeper understanding of the discourse is required, extended by meta-cognitive information available from multiple sources as self-explanations. In this context, we use a polyphonic model of discourse derived from Bakhtin's work as a paradigm for analyzing CSCL conversations, as well as cohesion graph building designed for creating an underlying discourse structure. This enables us to address both general texts and conversations and to incorporate comprehension and collaboration specific activities in a unique framework. As specificity of the analysis, in terms of *individual learning* we have focused on the identification of reading strategies and on providing a multi-dimensional textual complexity model integrating surface, word specific, morphology, syntax and semantic factors. Complementarily, the *collaborative learning* dimension is centered on the evaluation of participants' involvement, as well as on collaboration assessment through the use of two computational models: a *polyphonic model*, defined in terms of voice inter-animation, and a specific *social knowledge-building model*, derived from the specially designed cohesion graph corroborated with a proposed utterance scoring mechanism.

Our approach integrates advanced Natural Language Processing techniques and is focused on providing a qualitative estimation of the learning process. Therefore, two tightly coupled perspectives are taken into consideration: *comprehension* on one hand is centered on knowledge-building, self-explanations from which multiple reading strategies can be identified, whereas *collaboration*, on the other, can be seen as social involvement, ideas or voices generation, intertwining and inter-animation in a given context. Various *cognitive validations* for all our automated evaluation systems have been conducted and scenarios including the use of *ReaderBench*, our most advanced system, in different educational contexts have been built.

One of the most important goals of our model is to enhance understanding as a "mediator of learning" by providing automated feedback to both learners and teachers or tutors. The main benefits are its flexibility, extensibility and nevertheless specificity for covering multiple stages, starting from reading classroom materials, to discussing on specific topics in a collaborative manner, and finishing the feedback loop by verbalizing metacognitive thoughts in order to obtain a clear perspective over one's comprehension level and appropriate feedback about the collaborative learning processes.

Résumé

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 pour analyser les conversations collaboratives, et une approche théorique visant à unifier les activités de compréhension et de collaboration dans un même cadre, utilisant la construction de 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 le 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 interreliées 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. Des validations cognitives de nos différents systèmes d'évaluation automatique ont été réalisées, et nous avons conçu des scénarios d'utilisation de *ReaderBench*, notre système le plus avancé, dans différents contextes d'enseignement.

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.

Acknowledgements

The 'magic' of this thesis consists of a juxtaposition and synergy of multiple points of view that can be perceived as concentric circles in terms of research and academics, over a personal background of inter-relations. Moreover, as the generated network is so dense, I feel compelled to say that this thesis and all underlying activities are actually the result of teamwork and close collaboration with a lot of people, in various contexts.

Foremost, right in the center of this network are my mentors and supervisors, *Ștefan Trăușan-Matu* and *Philippe Dessus*, to whom I am extremely grateful for their overall guidance and, moreover, for the pleasure of working side-by-side in so many joint activities.

Afterwards comes a close circle of collaborators who provided external guidance and valuable feedback: *Maryse Bianco* and *Aurélie Nardy* with whom we closely worked on the design and validation of our latest system, *ReaderBench*; *Traian Rebedea* for setting the foundations of *PolyCAFe*'s implementation and for all the interesting talks on technical approaches; *Nicolae Nistor* for expanding our initial research towards the assessment of virtual Communities of Practice; *Ciprian Dobre* for effectively combining Natural Language Processing with distributed computing; *Nicolas Balacheff* for his orientation and overview of Technology-Enhanced Learning; *Carlo Giovannella* for the induced perspective in terms of Learning Analytics; *Laurent Besacier* for his excellent ideas in terms of the computational analysis of voice inter-animation; *Sonia Mandin* for providing us extremely useful annotated corpora and analysis metrics; *Thomas François* for the valuable inter-change of ideas in terms of textual complexity; *Danielle McNamara* for all her immediate responses and guidance towards improving our approach in terms of reading strategies and textual complexity; and *Mathieu Loiseau* for his overall support and feedback. A 'special' category that has not been integrated in the previous list consists of the jury members, *Adina Magda Florea*, *Costin Pribeanu*, *Stefano Cerri* and *Bruno de Lièvre* – with most of them we already had the pleasure to work together in various contexts, ranging from conferences to joint projects.

The following level, although the boundaries fade, consists of the colleagues from University Politehnica of Bucharest, *Vlad Posea* and *Costin Chiru*, the students directly involved in research with whom we have already published multiple articles (*Diana Lupan*, *Bogdan Oprescu, Corina Ciubuc*, *Mihnea Donciu* and *Mădălina Ioniță*), but also the colleagues from LSE, Grenoble, including other PhD students, for integration and an overall excellent atmosphere.

Now we shift the perspective towards the personal background, from which three points of view emerge: *Călin Tatomir* who especially encouraged me in terms of personal development and changed my overall perspective as a merger with pragmatic business concepts, *Silviu Hotăran*, my 'advocate'

for finishing this thesis, and *Varujan Pambuccian*, who made a strong point in following the research path. Last, but under no circumstances least, I am extremely grateful to my family and friends for all their support and understanding.

Also, as financial support and as larger context, the research presented in this thesis was partially supported by the 264207 ERRIC – Empowering Romanian Research on Intelligent Information Technologies/FP7-REGPOT-2010-1 project, the FP7 2008-212578 LTfLL – Language Technologies for Lifelong Learning project, the POSDRU/107/1.5/S/76909 Harnessing human capital in research through doctoral scholarships (ValueDoc) project and by an ANR (Agence Nationale de la Recherche) DEVCOMP – *Développement de Compétences* – grant.

List of Abbreviations

A.S.A.P.	Advanced System for Assessing Chat Participants
CAF	Complexity, Accuracy and Fluency
CBLE	Computer-based Learning Environment
CEFR	Common European Framework of Reference
Ch.A.M.P.	Chat Assessment and Modeling Program
CoI	Community of Inquiry
CoP	Community of Practice
CSCL	Computer Supported Collaborative Learning
DEVCOMP	Développement de Compétences
DRP	Degree of Reading Power
EA/AA	Exact/Adjacent Agreement
ERRIC	Empowering Romanian Research on Intelligent Information Technologies
FFL	French as Foreign Language
GA	Genetic Algorithm
HTML	HyperText Markup Language
ICC	Intra-Class Correlations
ICT	Information and Communications Technology
IR	Information Retrieval
KAMA	Kaufman Adaptive Moving Average
KB	Knowledge-Building
KSV	*Knowledge Space Visualizer*
LA	Learning Analytics
LD	Learning Design

LDA	Latent Dirichlet Allocation
LMS	Learning Management System
LSA	Latent Semantic Analysis
LTfLL	Language Technologies for Lifelong Learning
MOOC	Massively Online Open Courses
NLP	Natural Language Processing
PLE	Personal Learning Environment
PMI	Pointwise Mutual Information
PolyCAFe	Polyphony-based system for Collaboration Analysis and Feedback generation
POS	Part of Speech
RB	*ReaderBench*
RST	Rhetorical Structure Theory
SNA	Social Network Analysis
SRL	Self-Regulated Learning
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TASA	Touchstone Applied Science Associates, Inc.
TEL	Technology-Enhanced Learning
Tf-Idf	Term frequency – Inverse document frequency
XML	Extensible Markup Language
vCoP	Virtual Community of Practice
VMT	Virtual Math Teams
WOLF	WordNet Libre du Français
ZPD	Zone of Proximal Development

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Goals and Interests

In every instructional situation, reading textual materials and writing down thoughts are the core activities that represent both causes (from learner's viewpoint) and indicators of learning (from teacher's viewpoint). Reading is a cognitive activity whose oral or written traces are usually analyzed by teachers in order to infer either learners' comprehension or reading strategies. Hence reading and writing are core activities that every teacher has to assess on a daily basis. Reading materials have to be scaled or tailored to suit pupils' actual level, and reading strategies have to be analyzed for inferring learners' level of text processing and understanding. In conjunction, we must also consider the social dimension derived from the interaction with other learners, as well as the tutors, mediated by multiple discussion channels. In addition, the availability of Information and Communications Technologies (ICT) and the huge learning needs at a global level have induced the emergence of new communication scenarios such as those of Computer Supported Collaborative Learning (CSCL) that provide an alternative to classic learning scenarios, with emphasis on participation and collaboration that reflect comprehension, as well. This alternative does not address solely the "distance" between learners, but also the immediate availability of the whole set of resources on the Web, both passive (data) and active (humans).

From a different point of view, teacher's support of learners' reading and writing is difficult to be carried out on a large scale, therefore s/he should take care of a small number of students. Moreover, assessing textual materials and verbalizations is a cognitively demanding and subjectivity-laden activity. In addition, while regarding the collaborative learning perspective, the assessment of multi-participant conversations is also a time-consuming process, whose evaluation is cumbered by the intertwining of multiple discussion threads.

In this context, we considered it useful to provide a *unified vision of predicting and assessing comprehension* in order to *support individual* and *collaborative learning*. Whereas prediction is centered on an a priori evaluation of learning materials in terms of textual complexity, assessment is focused on an a posteriori processing of general texts and of learner productions (even conversation interventions). Moreover, the comprehension level can be predicted from texts also by considering cohesion and coherence that are required for obtaining a cognitive mental representation of the underlying discourse.

While considering the previous reading and writing loops that can be used to reflect learner comprehension, we can also make a clearer demarcation presented in Table 1 in terms of individual learning, collaborative learning and assessment. By considering these valences and the emphasis on specific learner or tutor educational activities, we have obtained the general context of performing all subsequent analyses.

Table 1. Read/write learning activities centered on comprehension.

	Read	Write
Understand (individual learning)	Text materials	Verbalizations of understanding, summary
Discuss (collaborative learning)	Utterances of peers in chat or forum discussions	Personal chat or forum utterances
Assess (tutor perspective)	Textual complexity Peers' contribution and involvement	Reading strategies

Starting from the previous context, we thus need appropriate tools to support the activities of both learners and of teachers, in various educational scenarios. Nevertheless, besides the actual automatization of specific tasks, we should also strive to ensure a cognitive meaning of all underlying processes and estimations. Therefore, empirical validations should be performed in order to ensure the traceability to human representations and the adequacy of the conducted experiments. This would enable the support of high-level learner and tutor cognitive activities through technology, as well as the possibility to perform automatic analyses for achieving better comprehension through reading strategies and for better collaborating, reflected in the social knowledge-build process.

Moreover, of particular interest is the interdisciplinary approach and domains intersection within our research. On one hand, we have *informatics* as support for building the tools to perform the

automatic assessments. Especially Natural Language Processing is also a determinant element within the analysis as it is required for obtaining a deeper representation and understanding of discourse. On the other hand, we have *educational psychology*, focused on precise validations, on the similarity to the human annotation processes and on comprehension modeling in terms of cohesion, coherence and textual complexity. In the end, we can also add a dialogical *philosophical framing*, centered on the polyphony and voice inter-animation.

1.2 Thesis Outline

As within our analysis we enforce a polyphonic model derived from dialogism, we prefer to present the outline of the thesis as an intertwining of 'voices'. Additionally, as a personal target, our aim was to induce balance within the internal structure of each chapter, as well as between different sections of the thesis. Therefore, the major 'forces' are the theoretical aspects on one hand and the experimental studies, on the other, whereas the 'voices' are represented by individual learning (see chapter *2*), collaborative learning (see chapter *3*), together with a computational perspective centered on discourse analysis (see chapter *4*). Moreover, we opted to have three theoretical chapters, each with three sections to increase the stability of the theoretical framing of the thesis. Although each chapter has a clear focus, links between sections have been established in order to highlight the inter-dependencies or the intertwining of the major directions: *2.1* Coherence and Comprehension is echoed in *4.1* Measures of Cohesion and Local Coherence from a computation perspective; *3.3* Metacognition and Self-regulation in CSCL creates an echo of *2.3* Reading Strategies in terms of self-regulated learning; key concepts from *3.1.2* Bakhtin's Dialogism as a Framework for CSCL are used for describing *4.2* Discourse Analysis and the Polyphonic Model; *4.3* Natural Language Processing Techniques and *3.2* Social Network Analysis represent together the main tools used in the empirical studies in order to perform the automatic analyses.

All the previous voices are later on reflected in the empirical part of the thesis that describes in chronological order the developed systems – *A.S.A.P.* (see *5.1 A.S.A.P.* – Advanced System for Assessing Chat Participants), *Ch.A.M.P.* (see *5.2 Ch.A.M.P.* – Chat Assessment and Modeling Program), *PolyCAFe* (see *6 PolyCAFe* – Polyphonic Conversation Analysis and Feedback) and *ReaderBench* (see *7 ReaderBench* (1) – Cohesion-based Discourse Analysis and Dialogism, *8 ReaderBench* (2) – Individual Assessment through Reading Strategies and Textual Complexity and *9 ReaderBench* (3) – Involvement and Collaboration Assessment through Cohesion and Dialogism).

As the presentation of the previous systems' facilities and of their corresponding validations is comprehension oriented, chapter *10* Discussion augments the educational perspective by including a detailed description of envisioned education scenarios, besides the presentation of advantages and faced problems. In the end, chapter *11* Conclusion briefly presents the contributions and major findings of this thesis, accompanied by further research directions.

Part 1 – Theoretical Aspects

Overview of Theoretical Aspects

In a nutshell, our goal is to support the comprehension processes in both individual and collaborative learning or, more specific, to support underlying personal and social knowledge-building processes, through the use of automatic tools, assessing notably textual cohesion of both inputs (read texts) and outputs (learners' productions). All these central concepts are aggregated within Figure 1 that highlights three levels of relationships, each one impacting the others.

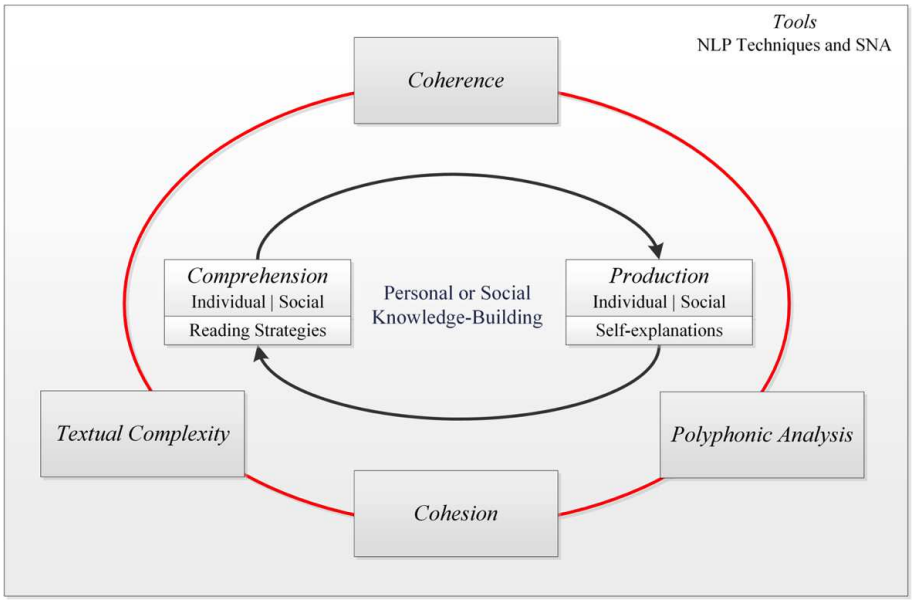


Figure 1. Integrated view of theoretical aspects and concepts.

Firstly, the *inner cycle* models the learning process from the knowledge-building perspective (Bereiter, 2002; Scardamalia, 2002; Stahl, 2006b), with both personal/individual and social/collaborative valences. This split can be also observed in terms of individual and social comprehension and productions, whereas our analysis addresses all types of dialogue, general texts, as well as conversations. Self-explanations (Chi, de Leeuw, Chui, & Lavancher, 1994; McNamara, 2004) can be considered particular cases of productions and can be used to evaluate comprehension through reading strategies (Millis & Magliano, 2012). Moreover, this circle highlights the main points of

interest – *comprehension*, as an overall aim that needs to be modeled, supported and achieved, and *productions*, as the elements of the analysis, both present within individual and collaborative learning scenarios.

Secondly, the *outer circle* consisting of cohesion, coherence, textual complexity and polyphonic analysis embodies the *assessment* process in terms of evaluating comprehension based on learner productions. The *polyphonic analysis* (Trausan-Matu, Stahl, & Sarmiento, 2006) uses the dialogism perspective (Bakhtin, 1981) for modeling the discourse and shares with the coherence concept the underlying target to achieve sense-making (Linell, 2009). Moreover, the polyphonic analysis provides relevant insight towards cohesion and textual complexity, including an informational view of local coherence, as voices' echoes can be seen as cohesive links spanning through the discourse. In addition, with regards to the situation model (van Dijk & Kintsch, 1983), *cohesion* and *coherence* play important roles while building a coherent mental representation.

The locations of each concept in Figure 1 were specifically chosen to reflect the most significant relations: 1/ *textual complexity* is related to cohesion and coherence and can directly impact comprehension with regards to the leaner's zone of proximal development (ZPD) (Vygotsky, 1978); 2/ *polyphonic analysis* highlights cohesive links through voice echoes and coherence through the merger of different points of view or through the inter-animation of voices, all being closely related to the learners' productions; 3/ *cohesion* can be considered the center of the proposed cohesion-based discourse model (see *7.2* Cohesion-based Discourse Analysis) and impacts both the polyphonic model and the text's complexity, as the lack of cohesion might artificially increase the perceived complexity; and 4/ in order to achieve a *coherent* representation, connectedness of discourse elements is essential, modeled through cohesion and the polyphonic model, whereas the textual complexity of the learner materials must be in the range that will challenge him/her, without causing frustration or the loss of motivation; in other words, *coherence* relies upon cohesion, appropriate textual complexity and a polyphonic weaving of voices.

Thirdly, the *background* consists of the tools, with emphasis on Natural Language Processing (NLP) techniques, necessary for performing discourse analysis, and Social Network Analysis, centered on interaction analysis. These tools are used for performing all computations and can be regarded as support for all underlying processes implemented within the presented systems (either the states of

art, or the developed systems – *A.S.A.P.*, *Ch.A.M.P*, *PolyCAFe* and *ReaderBench* – presented in detail in the empirical part of the thesis).

In the end, the goal was also to obtain two intersecting axes, each dominated by one of the directions of our inter-disciplinary research: one axis is oriented on the *cognitive psychology* perceptions of cohesion and coherence, whereas the other is more *informatics* oriented, with emphasis on discourse analysis and computational textual complexity. Moreover, comprehension used to support individual or collaborative learning and measured through the learner's productions is at the center of the diagram, as it can be represented from the learners' productions as an overlap of cohesion, coherence, textual complexity and polyphony.

In order to provide a clearer perspective of the theoretical aspects, Table 2 provides a mapping between each concept from Figure 1 and the corresponding theoretical section of the thesis.

Table 2. Mapping between key theoretical concepts and corresponding detailed descriptions.

Concept	Section with detailed description
Coherence and cohesion	*2.1.1* Coherence and Cohesion *4.1* Measures of Cohesion and Local Coherence
Comprehension	*2.1.2* Coherence and Comprehension
Textual Complexity	*2.1.3* Cohesion and Coherence versus Textual Complexity *2.2* Textual Complexity
Polyphonic Analysis	*3.1.2* Bakhtin's Dialogism as a Framework for CSCL *4.2* Discourse Analysis and the Polyphonic Model
Reading strategies and self-explanation	*2.3* Reading Strategies *3.3* Metacognition and Self-regulation in CSCL
Tools – Social Network Analysis	*3.2* Social Network Analysis
Tools – Natural Language Processing techniques	*4.3* Natural Language Processing Techniques

2 Individual Learning

This chapter addresses individual learning by firstly considering the multiple facets of cohesion and coherence, their links to comprehension and textual complexity, also grounding a computational view to be discussed later (see *4.1* Measures of Cohesion and Local Coherence). Afterwards, in tight relation to Figure 1 from the Overview of Theoretical Aspects, textual complexity is regarded from a computational perspective, highlighting multiple approaches. Later on, self-explanations, seen as specific learner productions, are used to support the learning process by making it more efficient and more focused on comprehension (McNamara, 2004).

As further implications, the principles and approaches presented in this chapter represent the foundations for the empirical studies centered on individual learning assessment (see *8.1* Identification of Reading Strategies and *8.2* Textual Complexity Analysis Model). In addition, in order to augment the learning perspective, the proposed educational scenario (see Figure 67 from section *10.3.1*) highlights: 1/ the *reading loop*, enriched by the document visualization facility – see Figure 42, 2/ the *writing loop* in which reading strategies are automatically identified from learner's self-explanations, 3/ the *gist loop*, which is more dynamic as learners produce keywords or select main sentences from the reading materials, as well as 4/ the *possibility* for tutors to *select* appropriate *textual materials* according to the learners' level based on their complexity level. Moreover, individual learning is also reflected in the pedagogical scenarios involving our system's transferability (see Table 38 from section *10.3.3*).

2.1 Coherence and Comprehension

2.1.1 Coherence and Cohesion

Cohesion and coherence are two central elements in linguistics that model a text's continuity. Although multiple definitions were given in time and the valences in terms of the differences between

the two concepts are greatly ranging from inter-changeability, mutual implication to partial independence, we opted for presenting the concepts in a view consistent with the computational perspective (see *4.1* Measures of Cohesion and Local Coherence). Nevertheless, cohesion and coherence are present in a well-written text characterized by unity and connectedness, in which individual sentences "hang" together and relate to one another (Celce-Murcia & Olshtain, 2000). *Cohesion* was introduced by Halliday and Hasan (1976) in terms of the cohesive ties or links between sentences of the same paragraph:

> "The concept of cohesion is a semantic one; it refers to relations of meaning that exist within the text, and that define it as a text. Cohesion occurs where the INTERPRETATION of some element in the discourse is dependent on that of another. The one PRESUPPOSES the other, in the sense that it cannot be effectively decoded except by recourse to it." (Halliday & Hasan, 1976, p. 4)

In addition, a text's structure is defined in terms of linguistic or semantic features contributing to its overall unity, in which cohesion is used for establishing the underlying structure of meaning: "cohesion does not concern what a text means; it concerns how the text is constructed as a semantic edifice" (Halliday & Hasan, 1976, p. 26). In addition, the concept of 'texture' represents more than cohesion, starts from the text's structure seen as an internal representation and introduces the "'macro-structure' of a text, that establishes it as a text of a particular kind – conversation, narrative, lyric, commercial correspondence and so on" (Halliday & Hasan, 1976, p. 324). Therefore cohesion addresses the local connections in a text based primarily on features that signal relationships between constituent elements (words or sentences).

On the other hand, *coherence* can be regarded as a connection between utterances used to "jointly integrate forms, meanings, and actions to make overall sense of what is said" (Schiffrin, 1987, p. 39). Coherence plays a central position in terms of text's sense-making and in communication, while addressing a deeper function of discourse, as it was seen by Widdowson (1978) – a relationship of illocutionary acts. In other words, coherence may be considered a "semantic property of discourses, based on the interpretation of each individual sentence relative to the interpretation of other sentences" (van Dijk, 1977, p. 93). Moreover, coherence between sentences is "based not only on the sequential relation between expressed and interpolated propositions, but also on the topic of discourse of a particular passage" (*ibid.*).

In addition, there are two levels of coherence that include *local coherence*, the linear or sequential relations between neighboring textual segments, and the *global* or *overall coherence* of discourse, mediated by the global theme of the document or the hierarchical topic progression (McNamara, Kintsch, Butler Songer, & Kintsch, 1996; Storrer, 2002). Unlike cohesion, coherence can be perceived as a global property of the text and may not be explicitly encoded within it (Beene, 1988). Additionally, coherence can be perceived from a psychologically-oriented point of view as a generalization of cohesion due to its multiple additional perspectives as the interaction with the reader's skill level, background knowledge and motivation that help forming the situation model (Tapiero, 2007). Similarly, coherence can be perceived as a "characteristic of the reader's mental representation of the text content" (Graesser, McNamara, Louwerse, & Cai, 2004, p. 193).

In terms of inter-dependencies between the two concepts, the main causality relation between them is the following: cohesion does not implicitly lead to coherence and cohesion by itself is insufficient for making a text coherent, as additional underlying linguistic properties are required to enable a truly coherent discourse. In other words, cohesion contributes to coherence (Hasan, 1984), but it is insufficient by itself to achieve it. Additionally, semantic connections between textual elements were classified by Enkvist (1987) into two types: 1/ connections through cohesion at surface level and 2/ connections through coherence at a more profound level, as sentences in a coherent text must "conform to the picture of one possible world in the experience or imagination of the receiver" (Enkvist, 1987, p. 126). In extent, coherence can be viewed as "the quality that makes a text conform to a consistent world picture and is therefore summarisable and interpretable" (Enkvist, 1990, p. 14).

Following this demarcation of connection types and aligned towards a more computational-oriented perspective, cohesion is centered on semantics as it depends on linguistic expressions and relies on the links between textual elements. On the other hand, coherence is clearly a pragmatic notion as it is focused on meaning, involves logical thinking and is highly dependent on external factors as one's knowledge or motivation. These later perspectives will be detailed in *4.1* Measures of Cohesion and Local Coherence. Starting from the definitions and inter-dependencies between cohesion and coherence, the next section goes one step further towards linking the concepts with the reader's comprehension and corresponding mental representations. Afterwards, the effect of cohesion and coherence in terms of influencing the perceived textual complexity is also analyzed.

2.1.2 Coherence and Comprehension

First and foremost, the understanding of a text requires building a coherent mental representation, commonly called a situation model (van Dijk & Kintsch, 1983). Moreover as initial studies reflect, referential cohesion diminishes the time required for reading and increases retention (de Villiers, 1974; Kintsch, Kozminsky, Streby, Mckoon, & Keenan, 1975). As cohesion is reflected in the connectivity in a text, the lack of appropriate connectives can substantially reduce the formation of inferences necessary for understanding a given text. While addressing coherence, deeper connections between concepts and ideas reduce the number of inferences needed to understand the text and to create a meaningful, globally coherent, mental representation, the coherence assumption: "The reader attempts to construct a meaning representation that is coherent at both local and global levels. Local coherence refers to structures and processes that organize elements, constituents, and referents of adjacent clauses or short sequences of clauses. Global coherence is established when local chunks of information are organized and interrelated into higher order chunks." (Graesser, Singer, & Trabasso, 1994, p. 371) Therefore, in a coherent text, relationships between different text elements can be easily observed (Zwaan & Singer, 2003) as comprehension depends on building these underlying relationships and connecting them to the situation model (Tapiero, 2007). Moreover, local coherence is a concept similar to cohesion, as in the following paragraphs we will focus on the coherence effects on readers' comprehension that can be partially considered cohesion effects.

On the other hand, incoherent texts generate loose and potentially wrong connections (cohesion level) and situation models (coherent representations level), and in the end possibly wrong decisions, as individuals might even quit making connections between text fragments if the text is very poorly composed. Nevertheless, in order to comprehend a text or to create its globally coherent cognitive representation, text coherence is only one of the three conditions that need to be met: "(a) the textual features support global coherence, (b) the reader has the prerequisite background knowledge, and (c) the reader does not have a specific goal that prevents understanding of the material" (Graesser et al., 1994, p. 378) (see Figure 2).

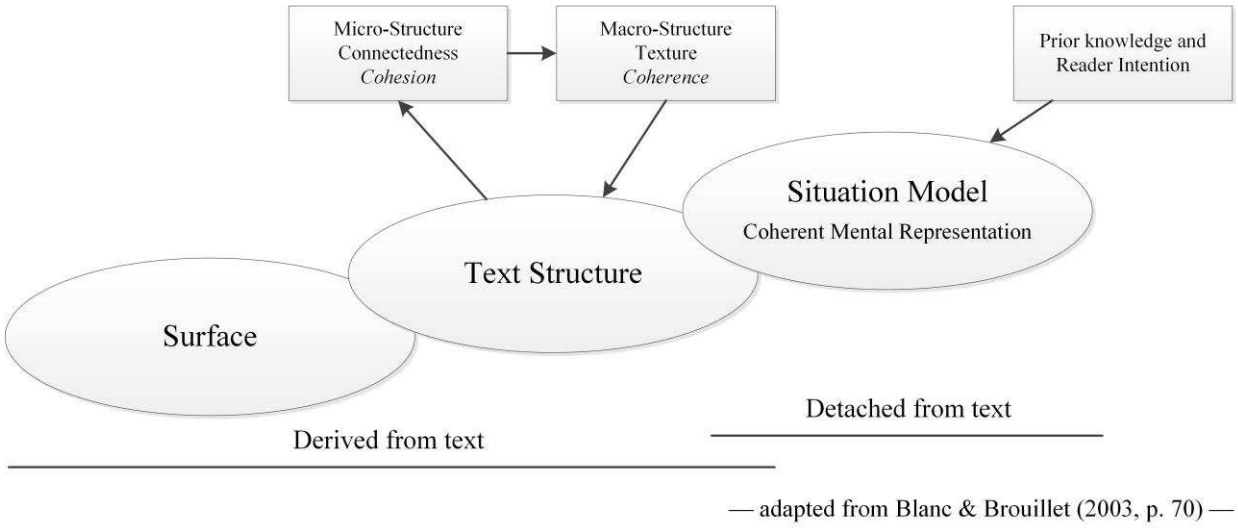


Figure 2. The three levels of comprehension representation.
The initial image was modified and augmented with additional concepts in order to best reflect our approach. The 'image' concept or the world's representation was removed, as it was not present within the conditions of a coherent cognitive representation (Graesser et al., 1994), whereas the micro and macro structures were complemented with cohesive/coherent perspectives. In order to comprehend a text, it becomes essential to create an initial overview, derived from the text, at both surface and deeper textual levels addressing cohesion and coherence, followed by the building of an internal mental representation, detached from the text – the *situational model* that is influenced by prior knowledge and intentions

As conclusions, the general trend is aligned with the following rule: the more coherent a text is, the more easily it can be understood. This assumption is nevertheless true for a general context, but when considering also the reader's knowledge level, it is true only for weak knowledgeable readers (McNamara et al., 1996). On the contrary, the effects generated by the presence of cohesion markers, the main elements computationally feasible to identify in terms of textual complexity, are mixed. The first effect that emerges addresses the decrease of retention rate for knowledgeable and/or good readers when the text's structure is too easy to reconstruct. Moreover, Alderson (2000) even suggests that the effects of cohesion on understanding and memory are rather low as readers are generally capable of making inferences, even in the absence of cohesive markers. More optimistic, Sanders and Noordman (2000) estimate that the presence of cohesion markers impacts reading, but does not affect the textual structure or meaning representation built by readers. On the other hand, Degand and Sanders (2002) sustain the opposite, as their experiments prove that readers presented with explicit discourse markers have better understood the initial text.

2.1.3 Cohesion and Coherence versus Textual Complexity

As cohesion can be regarded as the links that hold a text together and give it meaning, the mere use of semantically related words in a text does not directly correlate with textual complexity (see *2.2* Textual Complexity). For example:

(1) "John likes pancakes. The sky is blue. Your favorite cup is on the table."

(2) "John likes pancakes. He also likes cake. Cupcakes are John's favorites."

Both texts are rather simple in terms of words' general complexity and have comparable lengths and structure. However, the overall sequence of sentences in the first example lacks interconnections or cohesion. On the other hand, when analyzing the second example, there is a strong emerging point of view, as the text is about a person that enjoys cakes in all its forms. The repetitions of "likes" and "John" and the use of semantically related words ("pancakes", "cake", "cupcakes") ensure text cohesion. Therefore, cohesion in itself is not enough to distinguish texts in terms of complexity, but the lack of cohesion may increase textual complexity, as a text's proper understanding and representation become more difficult to achieve.

For coherence, things are clearer, as the intersection between a higher perspective of textual complexity and coherence lies in sense-making. Moreover, both concepts consider the learner's knowledge, motivation and interest. Therefore, the case of coherence is similar to cohesion (on which it actually relies), but with a tighter correlation to textual complexity, as coherence is more related to a meaningful mental representation of the discourse.

2.2 Textual Complexity

2.2.1 Overview of Textual Complexity

The idea of actually measuring and quantifying the complexity of texts has long been of interest for best aligning reading materials to the level of readers. Nevertheless, measuring textual complexity is in general a difficult task because the measure itself is relative to the reader and high differences in the perception for a given reading material can arise due to prior knowledge in the specific domain, familiarity with the language or to personal motivation and interest. Readability ease and comprehension are related to the readers' education, cognitive capabilities and background

experiences. Therefore a cognitive model of the reader must be taken into consideration and the measured complexity should be adapted to this model. Additionally, software implementing such functionalities should be adaptive in the sense that, for a given target audience, the estimated levels of textual complexity measured for specific texts should be adequate and relevant. In this context, textual complexity has a high impact on comprehension, retention and the zone of proximal development (ZPD) (Vygotsky, 1978) as this can be reflected in the range of learner materials that will challenge him/her, without causing the loss of motivation or frustration.

In addition, assessing the textual complexity of the material given to pupils is a common task that teachers encounter very often. However, this assessment cannot be performed without taking into account the actual pupils' reading proficiency and this point makes it time-consuming. Moreover, the impact of textual complexity on instruction and learning is important: pupils or students read faster and learn better if textual material is not too complex, nor too easy, as derived from ZPD (Vygotsky, 1978). All these points make the use of software that could calibrate the textual material according to the various reading levels appealing and useful.

From a pragmatic perspective, as considered by the Common Core State Standards Initiative (National Governors Association Center for Best Practices, 2010), textual complexity plays a leading role in evaluating student readiness for college and careers. In other words, the Standards' goal is to ensure that texts of steadily increasing complexity levels are presented to students so that textual complexity gaps can be reduced or eliminated in time. Therefore, a framework focused on three dimensions has been developed covering the three equally important perspectives of text complexity: quantitative, qualitative and a reader/task orientation. The most straightforward and computationally feasible measures of textual complexity are covered by quantitative factors, such as word frequency and sentence length. Secondly, qualitative dimensions of text complexity focus on the levels of meaning, structure, language conventionality and clarity, but also knowledge demands. Thirdly, reader and task considerations cover students' knowledge, motivation and interests.

Moreover, a key element when addressing textual complexity is the multitude of factors taken into consideration: "Without question the most important advances in readability research should result from the development of new linguistic variables." (Bormuth, 1966, p. 86). As the integrated textual complexity model within *ReaderBench* (Dascalu, Trausan-Matu, & Dessus, 2012; Dascalu, Dessus, Trausan-Matu, Bianco, & Nardy, in press) already incorporates the most popular and frequently

used readability formulas and factors, we opted for presenting them in detail in section *8.2* Textual Complexity Analysis Model, whereas the following section is focused on other computational approaches relevant to the task at hand.

2.2.2 Textual Complexity Computational Approaches

The first experiments for providing a comprehensive and automatic method of evaluating textual complexity were conducted in the research area of automatic essay grading, tightly connected to text complexity as essay grading can be seen as an assessment of complexity of the learning productions. One of the first and most popular systems is *E-Rater* (Burstein, Kaplan, Wolff, & Lu, 1996; Powers, Burstein, Chodorow, Fowles, & Kukich, 2001) that scores essays by extracting a set of features representing facets of writing quality. These features automatically extracted by *E-Rater*, later on applied in the assessment of textual complexity, include (Attali & Burstein, 2004): content analysis based on vocabulary measures, lexical complexity/diction, proportion of grammar, usage or mechanics errors, proportion of style comments, organization and development scores and features rewarding idiomatic phraseology. After evaluating each factor individually, all the previous features are combined in a statistical model to produce a final score estimate. When tested in Graduate Management Admission Test (GMAT), *E-Rater* scoring showed 87%-94% agreement (Chodorow & Burstein, 2004) with expert human reader (rates comparable to that between two expert readers who scored the same essays). Moreover, in order to emphasize the importance of cohesion and coherence in terms of essay grading that reflects the learner's textual complexity levels, *E-Rater* uses centering theory (Grosz, Weinstein, & Joshi, 1995) (see *4.2* Discourse Analysis and the Polyphonic Model) as a model to assess the complexity of inferences within the discourse by analyzing centers and centering transitions, and classifying the transitions into continue, retain, smooth-shift and rough-shift.

More sophisticated metrics have been implemented and widely adopted as solutions in various education programs (Nelson, Perfetti, Liben, & Liben, 2012): *Lexile* (MetaMetrics), *ATOS* (Renaissance Learning), Degrees of Reading Power: *DRP Analyzer* (Questar Assessment, Inc.), *REAP* (Carnegie Mellon University), *SourceRater* (Educational Testing Service), the *Pearson Reading Maturity Metric* (Pearson Knowledge Technologies), *Coh-Metrix* (University of Memphis) and *Dmesure* (Université Catholique de Louvain).

Individual Learning

The *Lexile Framework for Reading* (Stenner, 1996; Lennon & Burdick, 2004; Stenner, Koons, & Swartz, 2009) determines both the complexity of text and the individual's reading ability on the same developmental *Lexile* scale. This unification enables learners to search for targeted readings from a database of over 400 million entries (websites, books or articles). The main variables taken into account when computing the *Lexile* measure are word frequency and sentence length (Nelson et al., 2012). Moreover, an important categorization has been specified in terms of documents that should not be measured using the *Lexile Analyzer*, a categorization easily applicable to all automatic textual complexity assessment tools (http://www.lexile.com/tools/lexile-analyzer/step-1-what-texts-can-be-measured/), including our developed systems:

- *Student writing* – although textual complexity analysis is relevant for gaining insight of student's comprehension level, additional specific automatic essay grading factors or the identification of reading strategies should be used in conjunction with the complexity assessment (e.g., comparison to initial texts) in order to perform thorough evaluation.
- *Poetry* and *songs* – there are particular problems with phrase structure, ellipses, increased frequency of metaphors, epithets or enumerations; additionally, rhythm, rhyme and lyrics measures should also be considered.
- *Multiple-choice questions* – there is no sense of measuring the complexity of individual choices, taken one by one out of the context; moreover, low cohesion between the choices and elliptical grammatical structures might mislead the categorization.
- *Non-prose* – besides the previous categories, the connotations of images or pictures, although greatly impacting the comprehension of a material, are disregarded while performing an automatic assessment.
- *unconventionally punctuated or formatted text* – this generates problems at syntactic level while defining the dimensions of sentences, morphological while performing part-of-speech tagging and at pragmatic level as overall coherence can be greatly impacted by changes is punctuation.

In terms of prior text cleaning (applicable also for most automatic textual complexity assessment systems), "figures, tables, equations, titles, headings, footnotes/endnotes, numbered lists, non-standard characters, and pronunciation guides must be removed or altered manually" (Nelson et al.,

2012, p. 9). As misspellings can be detected automatically, it is also recommended to correct them by hand in order to improve accuracy.

ATOS (*Advantage-TASA Open Standard*) (Borman & Dowling, 2004) separates measurements formulas based on text target in two categories: *ATOS* for Text and *ATOS* for Books. Both formulas take into account the number of words per sentence, the average grade level of words (see below) and the number of characters per word. When addressing books, two additional important factors were used: the length of the book and variations in its internal structure (Renaissance Learning, 2011). The grade level of words is achieved via Graded Vocabulary List (Milone, 2012) consisting of about 24,000 words, developed using existing word lists, word frequency studies, vocabulary test results and linguistic experts. Tests were conducted on actual student book readings involving over 30,000 learners that passed through almost 1,000,000 books (Renaissance Learning, 2012b). *ATOS* is the default readability system integrated in the *Accelerated Reader* program used in approximately 50,000 schools in the U.S.A. (Nelson et al., 2012). This can be correlated to the fact that *ATOS* can be considered more reliable than *Lexile* when grading books as it is an open standard, extensive experiments were conducted for adjusting the graded vocabulary and *ATOS* can be considered more accurate, as it also enables learners to read a wider range of literature (Renaissance Learning, 2012a). In terms of text cleaning, *ATOS* is specific in contrast to other systems, as the user is not required to perform cleaning on the input text if its dimension is considerable; in this case, the complexity level emerges from the overall document.

Degrees of Reading Power (*DRP*) *Analyzer* (Koslin, Zeno, Koslin, Wainer, & Ivens, 1987; Zeno, Ivens, Millard, & Duvvuri, 1995) uses a derivation of a Bormuth mean cloze readability formula (Bormuth, 1966, 1969) based on three features: word length, sentence length and word familiarity (Nelson et al., 2012). The undergone experiments assume a group-administered cloze test consisting of selecting, from multiple-choice options, the correct word for deleted words in nonfiction paragraphs or passages on various topics (McNamara, Graesser, & Louwerse, in press). Text complexity is expressed in DRP with values ranging from 15- to 99+ and the score represents the learner reading level required to obtain 75 percent correct answers on the cloze tests. Moreover, the *TASA* (Touchstone Applied Science Associates, Inc.) corpus (approx. 13M words in 44k documents) contains texts already annotated with their corresponding DRP scores and could be divided into 12 categories of grade levels based on their *DRP* scores (McNamara et al., in press). Besides the traditional initial pre-processing in terms of text cleaning, the *DRP* score is determined for texts

ranging between 150 and 1000 words. The reliability for smaller texts is doubtful, whereas larger texts should be broken into segments, each analyzed individually.

REAder-specific Practice (*REAP*) Readability Tool (Heilman, Collins-Thompson, Callan, & Eskenazi, 2006; Dela Rosa & Eskenazi, 2011) is designed to teach students vocabulary through exposure to new words in a given context, dictated by the documents read by students. The goal is to derive the complexity level of each document from the individual complexities of the words it contains. Support vector machines (Cortes & Vapnik, 1995) and a simple bag-of-words approach (the order of the words in the text is not considered) are used for predicting the complexity class. *REAP* also provides a basic vocabulary difficulty estimate, whereas the used features include: word frequency, word length, sentence length and count, features of the sentences' or paragraphs' parse trees, and frequency of node elements (Nelson et al., 2012). As specificity, *REAP* removes words with less than 3 characters or function words.

SourceRater (Sheehan, Kostin, Futagi, & Flor, 2010) was designed to support teachers in evaluating complexity characteristics of used stimulus materials. In contrast to previous systems, *SourceRater* (Sheehan, Kostin, & Futagi, 2007) integrates a variety of NLP techniques to extract features like: syntactic complexity, vocabulary difficulty, level of abstractness, referential cohesion, connective cohesion, degree of academic orientation, degree of narrative orientation and paragraph structure. Complexity is estimated by enforcing one of the three separate regression models, each optimized for informational texts, literary texts or mixed ones. The main features used in assessing textual complexity include: word frequency, word length, word meaning features (concreteness, abstract, etc.), word syntactic features (tense, part of speech, proper names, negations, nominalizations, etc.), word types (academic verbs, academic word list), sentence and paragraph length, within-sentence or between-sentence cohesion measures, number of clauses (including type and depth) and text genre (Nelson et al., 2012). Of particular interest is the feedback module (Sheehan et al., 2010) useful for comparing individual text's complexity to a corpus with a known grade classification. More specifically, the module is focused on three dimensions that presume the identification of (Sheehan et al., 2010; Nelson et al., 2012): 1/ factors that might induce an unexpectedly low or high classification; 2/ segments of the initial text that might be more or less problematic to readers and 3/ overall text characteristics meaningful for technical review committees. In terms of cleaning, *SourceRater* requires proper paragraph markings and automatically detects non-standard characters, certain punctuation, and erroneous end-of-sentence markers (Nelson et al., 2012).

Word Maturity (Landauer, Kireyev, & Panaccione, 2011), introduced in the *Pearson Reading Maturity Metric*, uses language models built using Latent Semantic Analysis (LSA) (Landauer, Foltz, & Laham, 1998) (see *4.3.2* Semantic Similarity through Tagged LSA) to estimate how much language experience is required to achieve the adult meaning of each word, sentence and paragraph within a given text (Kireyev & Landauer, 2011). In extent, it models the degree to which a word is known at different levels of language exposure. Therefore, intermediate corpora of different complexity levels are used to train subsequent semantic LSA vector spaces reflecting intermediate level learner representations of concepts. After vector spaces are aligned using Procrustes Analysis (Krzanowski, 2000), word-meaning representations from each intermediate level are compared to the corresponding ones from the reference model. In the end, a word's maturity is reflected in its evolution throughout all subsequent spaces: simpler words are assimilated faster and their maturity scores reach the maximum value considerably faster than more elaborate words (see Figure 3). Pearson's *Reading Maturity Metric* also includes features like word length, sentence length, within sentence punctuation, sentence and paragraph complexity, order of information and semantic coherence (within and between sentences) (Nelson et al., 2012). Manual pre-processing requires the use of a consistent character encoding scheme (e.g., UTF-8) and the removal of non-textual elements (e.g., illustrations or equations).

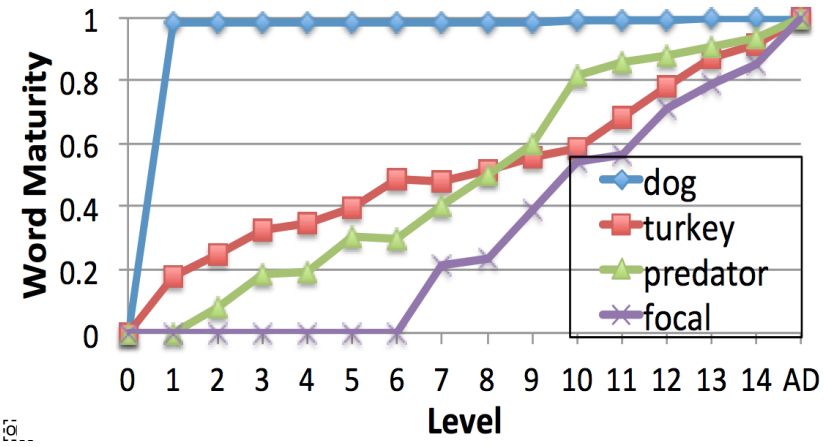


Figure 3. Word maturity curves for selected words (Kireyev & Landauer, 2011, p. 302). Simple words (e.g., "dog") obtain their adult meaning rather fast, while for specific words (e.g., "focal") it takes much longer to become known to any degree, moreover gain their final meaning (Kireyev & Landauer, 2011)

Coh-Metrix Text Easability Assessor (Graesser et al., 2004; McNamara, Louwerse, McCarthy, & Graesser, 2010) analyses the ease or difficulty of texts based on the following dimensions: narrative, syntactic simplicity, word concreteness, referential cohesion and deep cohesion. Whereas narrative

features measure whether a new passage is story-like and includes events and characters, syntactic simplicity addresses the complexity of the sentence syntax. On the other hand, word 'concreteness' classifies words into concrete and abstract. In the end, two types of cohesion are automatically computed: *referential cohesion* as overlap between sentences with respect to major words (nouns, verbs and adjectives) and to explicit ideas, and *deep cohesion* that evaluates meaning at deeper level, such as causal and temporal relations between events, actions, goals and states. Basic features used include: word frequency, word length, word meaning features (concreteness, number of senses, etc.) word syntactic features (part of speech, negations, etc.), sentence length, sentence complexity, paragraph length, within-sentence and between-sentence coherence and cohesion measures. Moreover, Latent Semantic Analysis (Landauer & Dumais, 1997) can be considered the central engine for determining the semantic overlap between sentences and overlaps. As a deeper insight, *Coh-Metrix*'s framework consists of five levels (Graesser, McNamara, & Kulikowich, 2011): 1/ words (covering word frequency, part-of-speech, number of senses, psychological ratings, Semantic content); 2/ syntax; 3/ the explicit text base (co-reference, lexical diversity, Latent Semantic Analysis.); 4/ the situation model; 5/ the discourse genre and rhetorical structure. Automatic preprocessing eliminates non-standard characters and certain types of punctuation, whereas the size of the analyzed texts, *Coh-Metrix* supports texts between 200 and 1,000 words (if larger, the process is applied sequentially on textual segments of fitting dimensions).

Dmesure (T. François, 2012; T. François & Miltsakaki, 2012) addresses lexical and syntactic complexity factors applied on French as a foreign language (FFL) texts. A multitude of factors were aggregated using different classifiers (e.g., multinomial logistic regression, decision trees, bagging and boosting, support vector machines) (X. Wu et al., 2008) in order to automatically generate language exercises fit to the learners' level. Texts and sentences were classified according to the CEFR scale (Common European Framework of Reference for Languages). Although the goal was to provide a comprehensive and extensible model for both English and French, extensive measurements were performed only for French as the training and evaluation corpus consisted of manually classified French documents (see section *8.3* for a detailed comparison between *Coh-Metrix*, respectively *Dmesure,* and our developed system – *ReaderBench*).

As an overview of all previously presented systems, most of them rely on surface factors that are not representative in terms of underlying cognitive processes for creating the situation model required for comprehension. Therefore, except for Pearson's *Reading Maturity Metric* and *Coh-Metrix*, there

is a low emphasis on semantic factors; moreover, solely Latent Semantic Analysis is used to model a text's cohesion, without taking into consideration other semantic similarity measures. From a different perspective, the dictionary-based approach from *ATOS*, although grounded by extensive measurements, is not easily extensible and requires constant updates as the language evolves. In addition, out of the presented systems, only *Dmesure* uses automatic classifiers for enhancing the prediction accuracy. In this context, our developed system, *ReaderBench* (see *8.2* Textual Complexity Analysis Model), integrates a multi-dimensional model based on Support Vector Machines comprising of the most frequently used textual complexity factors, with emphasis on semantics, and enhances the pre-processing step by applying a complete Natural Language Processing pipe (Manning & Schütze, 1999).

2.3 Reading Strategies

Moving from textual complexity to readers' comprehension assessment is not straightforward. Constructing textual coherence for readers requires that they are able to go beyond what is explicitly expressed. To do so, readers make use of cognitive procedures and processes, referred to as reading strategies, when those procedures are elicited through self-explanations (Millis & Magliano, 2012). Research on reading comprehension has shown that expert readers are strategic readers. They monitor their reading, being able to know at every moment their level of understanding. Moreover, when faced with a difficulty, learners can call upon regulation procedures, also called reading strategies (McNamara & Magliano, 2009). In this context, psychological and pedagogical research has revealed that people tend to understand better a text if they try to explain themselves what they have read (Chi et al., 1994; McNamara & Scott, 1999). Starting from these observations, techniques, such as *SERT* (Self-Explanation Reading Training) (McNamara, 2004, 2007), were developed to help students understand texts and to make the learning process more efficient and focused on comprehension. Reading strategies have been studied extensively with adolescent and adult readers using the think-aloud procedure that engages the reader to self-explain what they have understood so far, at specific breakpoints while reading, therefore providing insight in terms of comprehension. Moreover, self-regulation can be enhanced through the use of metacognitive reading strategies (Nash-Ditzel, 2010).

Four types of reading strategies are mainly used by expert readers (McNamara, 2004). *Paraphrasing* allows the reader to express what he/she understood from the explicit content of the text and can be

considered the first and essential step in the process of coherence building. *Text-based inferences*, for example causal and bridging strategies, build explicit relationships between two or more pieces of information in texts. On the other hand, *knowledge-based inferences* build relationships between the information in text and the reader's own knowledge and are essential to the situation model building process. *Control strategies* refer to the actual monitoring process when the reader is explicitly expressing what he/she has or has not understood. The diversity and richness of the strategies a reader carries out depend on many factors, either personal (proficiency, level of knowledge, motivation), or external (textual complexity).

Nevertheless, if we want students to be assisted while reading, one human expert can take care only after a small number of them, which makes it impossible for such training techniques to be used on a large scale. In this context, in many attempts to exploit MOOCs (Massively Online Open Courses) assistance is provided by peer students with the intrinsic risk of making mistakes. Moreover, assessing the content of a verbalization is a demanding and subjectivity-laden activity, which can be assisted by computer-based techniques. These are the main motives behind the idea of using a computer program instead of, or as support for, a human tutor.

Initial experiments were conducted by McNamara and her colleagues (O'Reilly, Sinclair, & McNamara, 2004) and *iSTART* (McNamara, Boonthum, & Levinstein, 2007; McNamara, O'Reilly, Rowe, Boonthum, & Levinstein, 2007) can be considered the first implemented system that addresses self-explanations (Jackson, Guess, & McNamara, 2009). It has various modules that explain the *SERT* method to the students, one that shows them how to use those techniques using a virtual student, and another training module that asks students to read texts and give verbalizations, evaluates them and provides an appropriate feedback. The main challenge raised by such a system is evaluating verbalizations given by pupils in accordance to the reading materials.

iSTART divides verbalizations into four main categories: irrelevant, paraphrases, verbalizations that use knowledge previously found in the text and verbalizations which use external knowledge from the students' experience. As stated in Landauer, McNamara, Dennis, and Kintsch (2007), it is easier to identify paraphrases and irrelevant explanations, but it is more difficult to identify and evaluate verbalizations that contain information coming from students' experience.

From a completely different point of view, Zhang et al. (2008) propose an alternative method of extracting explicit strategies for reading comprehension from spoken learner responses, as an

extension of the LISTEN's Reading Tutor project (https://www.cs.cmu.edu/~listen/). The approach used a logistic regression model on both synthetic and authentic data that consisted of tutorials, transcribed spoken responses, expert annotations and rationales for the recommendations in terms of features of the student responses. The educational scenario involved children that provided spoken responses, later on transcribed by hand, in a scripted instruction scenario by a reading expert. As inputs for the logistic regression model, annotations of each student's utterances are used containing information on how he/she should have replied to the prompt and why.

Of particular interest are the experiments performed by Nardy, Bianco, Toffa, Rémond, and Dessus (in press) within the ANR DEVCOMP project that emphasized the control and regulation of comprehension through reading strategies. The experiments were conducted on pupils (3rd – 5th grade, 8 – 11 years old) that read aloud two stories and were asked at predefined moments to self-explain their impressions and thoughts about the reading material. An adapted annotation methodology was devised starting from the coding scheme of McNamara (2004) that covered: paraphrases, textual inferences, knowledge inferences, self-evaluations and other (see Figure 4). The "other" category is very close to the irrelevant category of McNamara (2004), as it aggregates irrelevant, as well as not understandable statements.

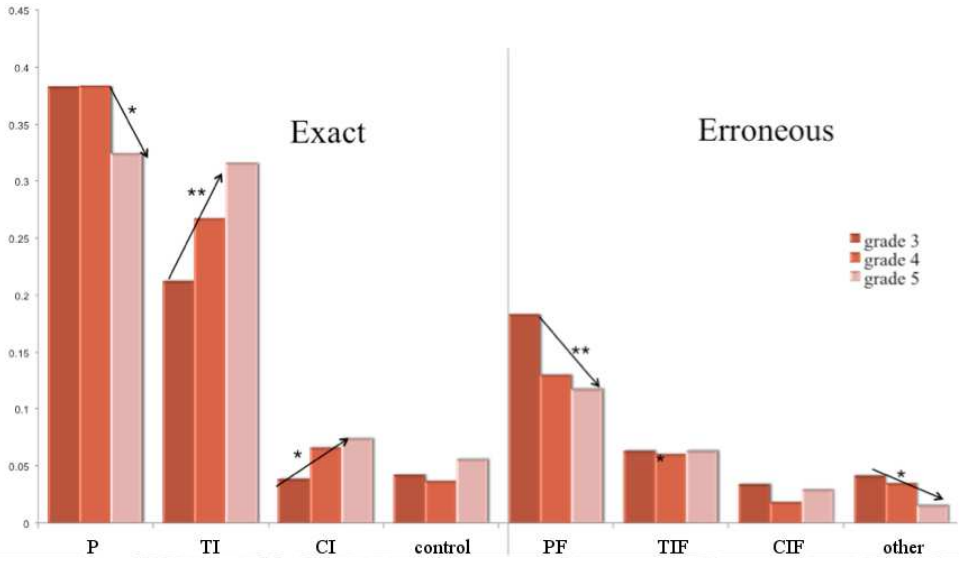


Figure 4. Mean frequencies of each type of strategies elicited by self-explanations from 3rd to 5th grades pupils (Nardy et al., in press).

Two dominant strategies were identified: paraphrases (P) and text-based inferences (TI); Text-based inferences (TI) frequency increases from grade 3 to 5; False Paraphrases (PF) frequency decreases from grade 3 to 5; Knowledge-based inferences (CI) remain rare, but their frequency doubles from grade 3 to 5

The conclusions of the performed study were remarkable: 1/ self-explanations are a useful tool to access reading strategies in young children (8 to 11 years old) that dispose of all the strategies described for older children; 2/ a link between the ability to paraphrase and to use text-based inferences, on one hand, and comprehension and extraction of internal text coherence traits, on the other, could be developed: a better comprehension in this age range is tied to less false paraphrases and more text-based inferences and 3/ age reduces the mediating effect of verbal logic ability for text-based inferences.

Starting from the previous experiments, one of the goals of *ReaderBench* (see *8.1* Identification of Reading Strategies) was to enable the usage of new texts with little or no human intervention, providing fully automatic assessment of reading strategies as support for human teachers (see also section *8.3* for a detailed comparison between *iSTART* and *ReaderBench*).

3 Collaborative Learning

This chapter creates a framing in term of collaborative learning, as it is focused on presenting *chats* that emerged as a viable alternative to the classic view of learning within Computer Supported Collaborative Learning (CSCL): Bakhtin's *dialogism* (Bakhtin, 1981) that defines the CSCL paradigms, *computational approaches* and *Social Network Analysis* (SNA) as the main Learning Analytics (LA) tool for modeling interaction between conversation participants. Moreover, a parallel is drawn between CSCL and individual learning through self-regulated learning in terms of metacognition (see *2.3* Reading Strategies), as learning strategies need to be effectively used in order to participate in collaborative interactions. The learner must monitor, regulate and control his/her cognition, motivation, behavior and emotions within the collaborative educational context. With regards to Figure 1 from the Overview of Theoretical Aspects, this chapter emphasizes the social dimension of knowledge-building; productions are mostly reflected in the interventions from CSCL environments, whereas comprehension obtains different facets, ranging from involvement, collaboration and self-regulation in CSCL.

In addition, the core dialogism concepts and the principles of rhythm analysis applied on the identified voices define the polyphonic model presented in *4.2* Discourse Analysis and the Polyphonic Model that is later on extended in *7.5* Dialogism and Voice Inter-Animation and in *9.2.2* Dialogical Voice Inter-Animation Model. Of particular interest is also the proposed educational scenario (see Figure 68 from section *10.3.1*) in which the writing loop emphasizes social knowledge-building, while the reading loop that takes full advantage of the designed chat visualization facility, augments personal knowledge-building. The collaborative learning dimension is also highlighted in the pedagogical scenarios involving our system's transferability (see Table 39 from section *10.3.3*).

3.1 Computer Supported Collaborative Learning

3.1.1 Chats as Support for Social Cognition

As the web evolved into a social environment, other communication channels were developed allowing users to exchange ideas, thoughts and information worldwide. Chat is probably the practical and most simple to use web communication technology at this point and thus justifies its popularity among users. Furthermore, during the last couple of years, various chat-like systems have appeared and have become very appealing: Twitter and Facebook status updates are just the most renowned examples. It seems that the dialog (which implies real-time, synchronous inter-change of utterances) using short textual messages fits naturally the needs of a large number of users.

Although most times chat-like technologies are used only for socialization and similar activities, they have also been adopted in education. In informal learning, discussions between the members of communities of practice often take place using chats. Nevertheless, instant messaging determined a change in the way collaborative work is regarded, becoming a viable alternative to the classic view of learning: Computer Supported Collaborative Learning (CSCL) (Stahl, 2006b) that advocates for the use of chats as a supplement for standard teaching and learning strategies (Stahl, 2006b). In this manner, chat has been introduced in formal education as well and is used by students to solve problems and debate difficult topics in order to develop their knowledge about a given domain and to learn from their peers.

Moreover, during the last years, several CSCL applications appeared that used different kinds of web communication and collaboration tools and environments like forums, chats, blogs, wikis etc. However, this situation raised new problems for tutors that needed to assess and provide feedback to the students participating in chat conversations related to a course, due to the high volume of information; thus, an automatic system's help would be required. For example, a professor's evaluation is an extremely time consuming process (Trausan-Matu, 2010a) and social networks and natural language processing would be helpful. Moreover, it is considerably more difficult to assess a collaborative chat than a normal text written by an individual student. Nevertheless, the development of such assessment tools is essentially difficult because of natural language processing, constituting one of the most intricate domains of artificial intelligence.

From a different perspective, CSCL is based on a totally different paradigm (Koschmann, 1999; Stahl, 2006b), grounded on dialogism and the social-cultural ideas of Vygotsky (1978) and Bakhtin (1981), which appeared decades before the invention of the computer. In contrast with the classical, cognitive approach to artificial intelligence, CSCL moves towards a socio-cultural paradigm, focusing on the idea that knowledge is created collaboratively through the process of social knowledge-building (Bereiter, 2002; Scardamalia, 2002; Stahl, 2006b) (see Figure 5).

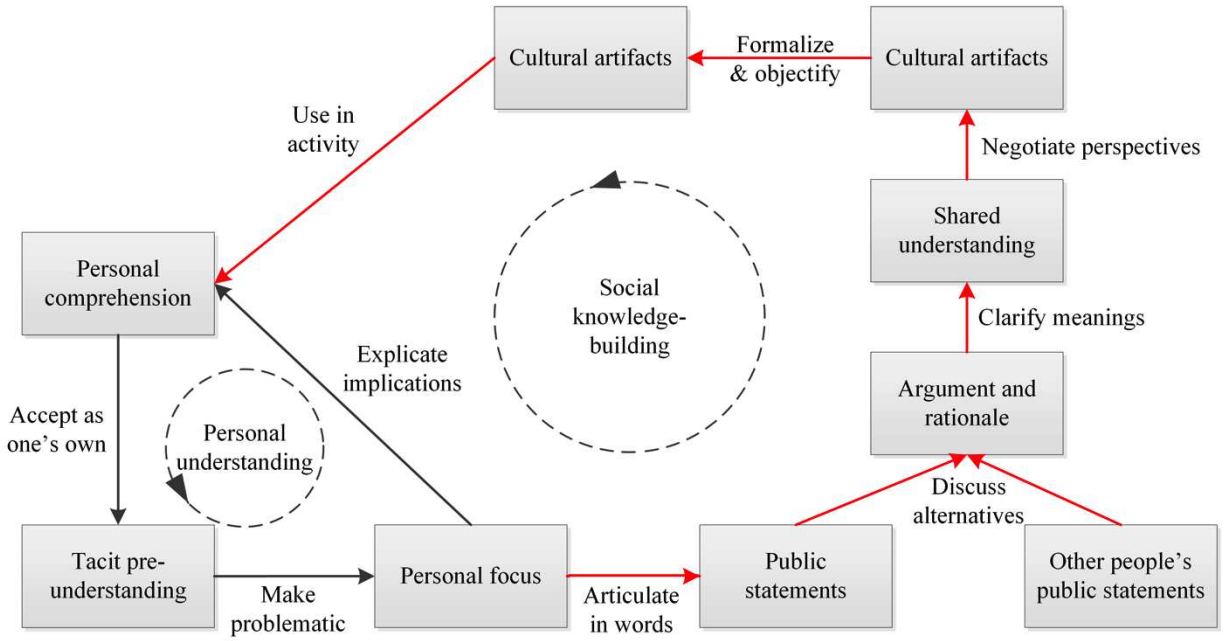


Figure 5. Diagram of knowledge-building processes, with emphasis on the social/collaborative dimension (after Stahl, 2006b, ch. 9).

Important phases in knowledge-building, with emphasis on the social component, are presented using the following conventions: arrows = transformative processes & rectangles = results of the processes, expressed as different forms of knowledge. The diagram presents the learning process as a mutual constitution of individual (personal understanding loop in black) and social knowledge-building (the red loop) (Lave & Wenger, 1991; Stahl, 2006b). Starting from individual personal beliefs, multiple transformations occur from a social perspective through the interaction with other people and with shared culture

However, few systems provide complex analysis and feedback facilities on chat and forum discussions in order to be useful for learners and tutors. There are at least two factors that explain this situation. The first factor is that, even if dialogism (Bakhtin, 1981) (see *3.1.2* Bakhtin's Dialogism), is considered as a fundamental paradigm of CSCL (Koschmann, 1999; Stahl, 2006b), extremely few software implementations started from it (Trausan-Matu, Rebedea, Dragan, & Alexandru, 2007; Trausan-Matu & Rebedea, 2010; Rebedea, Dascalu, Trausan-Matu, Armitt, & Chiru, 2011; Dascalu, Trausan-Matu, & Dessus, in press). The second factor is related to the fact that the majority of collaborations in CSCL are based on the exchange of text messages. Thus, another problem arises

from the fact that current Natural Language Processing (NLP) systems are far from providing reliable text understanding systems.

In addition, as the developed systems (*A.S.A.P.* – see Section *5.1*, *Ch.A.M.P.* – see Section *5.2*, *PolyCAFe* – see Chapter *6* and *ReaderBench* – see Chapter *9*) analyze in extent chat conversations, a deeper insight is more than appropriate. In general, in Internet chat systems (e.g. Yahoo Messenger in a conference style or similar environments), conversations are rather written, not spoken (and afterwards transcribed). This context allows more than two users to participate to the chat conversation. However, a problem is that multiple discussion threads may occur in parallel and, for example, if two participants write a question at a very short time interval, and a third answers, it may be very difficult to determine to whom it replied. A solution to this problem was spontaneously found: users refer the addressing person explicitly with the "@" sign in front: "@john you're right". Another solution was provided in the VMT environment (Stahl, 2009a), used in all our systems: for referencing, a user may click on the previous utterances (Holmer, Kienle, & Wessner, 2006) (see Figure 6). Moreover, referencing may be also extended to the provided whiteboard. Written chat conversations with environments that allow explicit referencing encourage the existence of parallel threads of discussion. More important, inter-animation processes appear among these threads, following patterns similar to those of counterpoint in polyphonic music (Trausan-Matu et al., 2006; Trausan-Matu & Rebedea, 2009).

Figure 6. Multiple discussion threads highlighted through the explicit referencing facility (Holmer et al., 2006).

From an educational perspective, a typical considered case is that of small virtual groups using chat systems for learning together (Stahl, 2006b; Trausan-Matu, Stahl, & Sarmiento, 2007). CSCL is a change of vision on learning replacing the idea of the transfer of knowledge from a human or a written source to the student. The new idea is that learning should empower the students to become participants in a discourse: "rather than speaking about 'acquisition of knowledge', many people prefer to view learning as becoming a participant in a certain discourse" (Sfard, 2000, p. 160). A natural consequence is that in order to provide automatic assessment and feedback generation, the system should be able to analyze students' discourse and therefore theories on discourse are needed (see *4.2* Discourse Analysis and the Polyphonic Model), corroborated with natural language techniques (see *4.3* Natural Language Processing Techniques).

3.1.2 Bakhtin's Dialogism as a Framework for CSCL

Dialogism was introduced by the Russian philosopher Mikhail Bakhtin (Bakhtin, 1981, 1984) and covers a broader, more abstract and comprehensive sense of dialogue that is reflected in "any kind of human sense-making, semiotic practice, action, interaction, thinking or communication, as long as these phenomena are 'dialogically' or 'dialogistically' understood" (Linell, 2009). This provides a differentiation criteria in terms of the classic dialogue theories that are focused on the interactions between two or more individuals, mutually present in real-time or with accepted delayed responses, using different communication channels (of particular interest here are the computer-supported "dialogues"). The dialogical framework is therefore centered on sense-making, with emphasis on (Linell, 2009): 1/ *action* as Wertsch (1998) suggests that the mind is constructed as actions and meaning is achieved through interaction with others and the world, in a given context; 2/ *cognition* as we acquire knowledge about the world and ascribe meaning to it through language and interaction, within a specific context; and 3/ *communication* that assumes the interaction with others generates the meaning of discourse and also incorporates a strong cognitive component as "every authentic function of the human spirit […] embodies an original, formative power" (Cassirer, 1953, p. 78).

In this context, other-orientation defined as the inter-relations with 'others' (that can embody an individual, as a concrete person, or a group, as a generalized perspective) plays a central role from which "*responsitivity* and *anticipation* in action and interaction are part and parcel of all pieces of discourse" (Linell, 2009, p. 13). Therefore, from a dialogic perspective, there are no 'autonomous' subjects who are isolated, who think, speak and act in and by themselves, as all actions, thoughts and

expressed utterances are dependent to the interaction with others, and their corresponding actions (Linell, 2009):

> "Every word is directed towards an *answer* and cannot escape the profound influence of the answering word that it anticipates. […] Responsive understanding is a fundamental force, one that anticipates in the foundation of discourse, and it is moreover an active understanding, one that discourse senses as resistance or support enriching the discourse." (Bakhtin, 1981, p. 280) "The word in language is half someone else's. It becomes 'one's own' only when the speaker populates it with his own intentions, his own accent […]. Prior to this moment of appropriation, the word does not exist in a neutral or impersonal language […], but rather exists in other people's mouths, in other people's concrete contexts, serving other people's intentions" (Bakhtin, 1981, pp. 293-294)

Moreover, dialogue can be also perceived within a broader and more abstract perspective as being capable of grasping multiple valences: 'internal dialogue within the self' or 'internal dialogue' (Linell, 2009, ch. 6), 'dialogical exploration of the environment' (Linell, 2009, ch. 7), 'dialogue with artifacts' (Linell, 2009, ch. 16), 'dialogue between ideas' (Marková, Linell, Grossen, & Salazar Orvig, 2007, ch. 6) or 'paradigms' (Linell, 2005, ch. 6). Nevertheless, in each context, discourse is modeled from a dialogical perspective as interaction with others, essential towards building meaning and understanding.

With regards to Computer Supported Collaborative Learning, dialogism was proposed by Koschmann (1999) as a paradigm for CSCL, its key features being multivocality and polyphony. Wegerif (2006) also considered dialogism as a theoretical starting point that can be used for developing tools to teach thinking skills. Moreover, Wegerif believes that inter-animation is a key component for the success of collaborative learning.

In order to properly introduce the polyphonic model presented in detail in *4.2* Discourse Analysis and the Polyphonic Model and later on used within *PolyCAFe* (see *6 PolyCAFe* – Polyphonic Conversation Analysis and Feedback) and *ReaderBench* (see *7.5* Dialogism and Voice Inter-Animation and *9 ReaderBench* (3) – Involvement and Collaboration Assessment through Cohesion and Dialogism), we must first present the three core and inter-dependent concepts of discourse analysis: *utterances* briefly defined as units of analysis, *voices* as distinctive points of view emerging from the ongoing discussion and *echoes* as the replication of a certain voice with further implications

in the discourse. Similar to some extent to the dialogical discourse analysis proposed by Linell (2001) and Marková et al. (2007) focused on the dynamics and recurrence of topics ('themes') and their rhetoric expressions (e.g., analogies, distinctions, metaphors, use of quotes) (Marková et al., 2007), all computational perspectives are inevitably limiting while analyzing the dialogical nature or discourse: "it is indeed impossible to be 'completely dialogical', if one wants to be systematic and contribute to a cumulative scientific endeavor" (Linell, 2009, p. 383). This also augments the duality between individual involvement and actual collaboration throughout a given CSCL conversation, as it is impossible to focus on both the animation of other participants and sustainably providing meaningful utterances; in the end, a balance needs to be achieved between individuals, without encouraging domination of the discourse in terms of participation.

A Utterance

Utterances can be defined as pieces of text whose boundaries are represented by the change of speech subject (Bakhtin, 1986) and embed the central unit of analysis of the discourse. Utterances express both acts of communication and pieces of discourse (Linell, 2009) and direct the path and evolution of the ongoing conversation in terms of future development. Although the complexity of an utterance may vary greatly from a simple word or interjection to a set of inter-twined utterances or even to an entire novel (Bakhtin, 1986), our analysis adheres to Dong's perspective of separating utterances based on turn-taking events between speakers (Dong, 2009). Therefore, introducing a new point of view or intervention from a different participant divides the discourse by changing the inner, ongoing perspective of the current speaker. At a more fine-grained level, words seen as the constituents of utterances provide the liaisons between utterances and deepen the perspective of other's interventions into one's discourse:

> "When we select words in the process of constructing an utterance, we by no means always take them from the system of language in their neutral, *dictionary* form. We usually take them from *other utterances*, and mainly from utterances that are kindred to our genre, that is, in theme, composition, or style". (Bakhtin, 1986, p. 87)

Moreover, listeners contribute to meaning, their responses can be considered a consequence of previously uttered elements, and this perpetual shift between the listener/speaker states models the dialogic perspective of the turn-taking exchange of utterances. Moreover, as building understanding

represents the common ground for all involved parties, elicitation can be considered the driving engine of a conversation:

> "Any understanding of live speech, of live utterance, is inherently responsive, although the degree of this activity varies extremely. Any understanding is imbued with response and necessarily elicits it in one form or another: the listener becomes the speaker. […] And the speaker himself is oriented precisely toward such an actively responsive understanding […] he expects response, agreement, sympathy, objection, execution, and so forth" (Bakhtin, 1986, pp. 68-69)

B Voice

A voice expresses a distinct position, a point of view, even an utterance or an event with further influence in the conversation. All preconditions are met by assuming that each utterance is read or heard, remembered and further discussed, therefore having an impact in the discourse (Trausan-Matu & Rebedea, 2009). Moreover, a voice may be expressed as a perspective on topics (Linell, 2009) of a singular participant or of a group sharing a similar insight on the topical domain. Therefore, opinions or perspectives on topics that are socially generated and sustained live in the "circulation of ideas" in conversations (F. François, 1993; Hudelot, 1994; Salazar Orvig, 1999). Individuals internalize and assimilate these ideas, later to re-emit them as voices that reflect their personal point of view; this can be also viewed as a "voting" process in terms of the uttered ideas, followed by an alignment to other individuals sharing the same perspective (Linell, 2009).

With regards to a single individual, he may adhere, personalize and express several different voices by interacting with other people based on his formal background, education and attitude towards the topic at hand. Therefore, besides internal voices embedding personal perspectives and external voices uttered by other individuals and expressing the influence of others on one's opinion, generalized voices emerge to which a larger group of people consent.

On the other hand, starting from the previously defined unit of analysis, an utterance may become a voice and reflect echoes and overtones of previous ones (Bakhtin, 1986). In this context, ventriloquism (Bakhtin, 1984), which is the (re)emitting of a voice by another one, gains larger implications in the sense that the entire discourse is governed by voice inter-animation and by the

influence of each voice or utterance measured in terms of its strength and further implications in subsequent utterances.

> "Dialogic orientation of discourse is a phenomenon that is, of course, a property of any discourse. On all its various routes towards the object, in all its directions, the word encounters an *alien* word, and cannot help encountering it in a living, tension-filled interaction. Only the mythical Adam […] could really have escaped from start to finish this dialogic inter-orientation with the alien word that occurs in the object. Concrete historical human discourse does not have this privilege: it can deviate from such inter-orientation only on a conditional basis and only to a certain degree." (Bakhtin, 1981, p. 127)

Obviously, utterances may contain more than a single voice, as well as alien voices to which the current voice refers to (Trausan-Matu & Stahl, 2007). Moreover, inter-animation considers two dimensions: *longitudinal* or chronologically sequential and *transversal*, following constraints similar to the counterpoint rules in music (Trausan-Matu & Stahl, 2007). Through transitivity, voices may accumulate during a conversation, generating either a joint/consonant discourse (Trausan-Matu & Stahl, 2007; Trausan-Matu & Rebedea, 2009) or dissonances between voices, inherently present, that need to be addressed (Trausan-Matu, in press). Therefore, as also Bakhtin (1981) noticed, individuals face both centrifugal (divergent, towards difference) and centripetal (convergent, towards unity) forces (Trausan-Matu, in press).

From a different point of view, in order to benefit mostly from collaboration, the main goal of a discussion can be defined in terms of voice inter-animation and achieving true *polyphony* (Bakhtin, 1984). Polyphony is closely related to the musical concept from which it was derived and encapsulates multiple points of view and voices. From Bakhtin's point of view, Dostoevsky's prose can be considered a true representation of polyphony because each character can be considered an individual voice, distinct from others. Moreover, Dostoevsky's work presents conflicting views, not just various angles and multiple perspectives, nor a single all-knowing and overwhelming vision, common among most writers; all these aspects should also be covered in a truly collaborative conversation. However, voices express ideas and opinions and by summing up multiple voices co-occurring within the same discussion thread or expressed by the same participant, 'poly-vocality' or polyphony can be used in order to perform a deep dialogical discourse analysis.

C Echo

A context is a slice of a discussion thread characterized by high internal cohesion and rather loose coupling with other parts of the conversation. A central voice emerges from a context, brings cognitive and creative significance and by its evolution in time models the unfinalized potential of that specific context (Bakhtin, 1986). The relation is bi-univocal in the sense that a context can encapsulate multiple voices and by merging all perspectives the context can be defined.

The echo of a specific voice represents its replication in time with enough strength to influence other voices in one or more contexts. Two types of echoes can be identified: individual ones, when a participant internalizes a voice, and collective echoes, when multiple participants react to a voice, enriching the context.

There is no predefined conclusion or ending to a context and by adding new voices and by considering echoes of previous ones to a context, the perspective might change. Snapshots of the current context can be taken, but there is no certain path of evolution, whereas echoes of current voices might influence subsequent utterances and model the internal voices of each participant.

After analyzing all core concepts, two major effects were identified and taken into consideration in our analysis. Firstly, a *retrospective*, synergic effect, based on overlapping voices from previous utterances and their corresponding echoes, influences the current utterance and the conversation context. Secondly, a *prospective* effect expresses further implications in discussion threads in terms of echoes and shapes the context, highlighting the unfinalizable, dialogic nature of the discussion.

> "Each utterance is filled with echoes and reverberations of other utterances to which it is related by the communality of the sphere of speech communication. Every utterance must be regarded primarily as a *response* to preceding utterances of the given sphere […]. Each utterance refutes, affirms, supplements, and relies on the others, presupposes them to be known, and somehow takes them into account." (Bakhtin, 1986, p. 91)

By combining all previous perspectives and key concepts, collaboration is generated through voice intertwining and inter-animation with other individuals. In other words, the dialogic nature of a discussion based on the inter-animation of voices can be used to evaluate the quality of the collaborative learning process, in tight relation with the actual interactions between the participants (see *9 ReaderBench* (3) – Involvement and Collaboration Assessment through Cohesion and

Dialogism). In addition, as a liaison to other theories that support CSCL, knowledge-building (Bereiter, 2002; Scardamalia, 2002; Stahl, 2006b) can be also perceived from a dialogic perspective as meaning and understanding that are generated from voice inter-animation.

D Rhythm Analysis

Of particular interest when discussing about polyphony in terms of dialogism is the comparison to music from which this concept was derived. Therefore, this section will focus on providing an empirical comparison of rhythm in language and music, as well as an extension towards collaboration and creativity. Overall, rather few attempts, mostly related to jazz, exist to compare music and language from the perspective of interactional influences among performers that would enhance group creativity (Sawyer, 1992; Berliner, 1994; Monson, 1996; Sawyer, 2003). As most approaches from structuralism focus on monophonic music, it clearly becomes "very difficult to apply the same procedure to polyphonic structures" (Ruwet, 1972, p. 116) in order to model, through music, the polyphonic interaction needed in group creativity (Sawyer, 2003).

Chafe (1997) addressed the "polyphony" of everyday conversations invoking a musical metaphor as "we need to avoid reifying our transcripts, keeping always in mind that they are an artificial freezing of phenomena which are in constant change" (Chafe, 1997, p. 52). In this context, conversations can be seen as a form of group creativity and can be perceived as an interactional dance, a polyphonic duet (Sawyer, 2003). Moreover, Sawyer (2003) proposed a model of group creativity derived from jazz ensembles and improvisational theater groups that was applied on a wide range of collaborating groups, including classrooms settings and innovative teams in organizations. The analogy of conversations to music is built around the moment-to-moment communication among jazz musicians and improvising actors that generates a result better than the sum of its parts, similar to collaboration.

From a different point of view, the study performed by Patel and Daniele (2003) compares rhythmic patterns in English and French language with classical music. Although significant differences were observed between English and French musical themes, as well as spoken language (see Figure 7), the conducted experiments can be considered an "empirical basis for the claim that spoken prosody leaves an imprint on the music of a culture" (Patel & Daniele, 2003, p. B35) as the similarities of the two generated graphics are striking. Moreover, London and Jones (2011) have performed a reanalysis of the data from Patel and Daniele (2003) after enforcing specific rhythmic refinements

for the application of the nPVI (normalized pairwise variability index) in musical contexts that included the use of 1/ higher levels of the rhythmic structure, 2/ a metrical structure (duple versus triple) and 3/ an alternative coding for surface durations. The previous studies deepen the similarity of perspectives between language and music, in terms of polyphony and rhythm analysis.

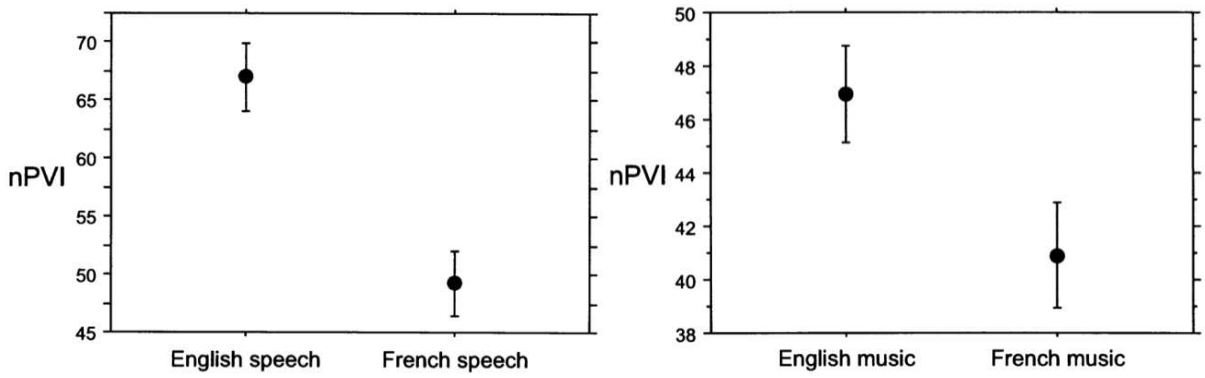


Figure 7. a. Linguistic nPVI (normalized pairwise variability index) values for sentences in British English and standard French versus b. Musical nPVI values for themes in English and French instrumental classical music (Patel & Daniele, 2003, pp. B38, B41).

More specific to CSCL, the perceived sense of sustained time and the rhythms of life rely upon the narratives people tell themselves (Bruner, 1990; Sarmiento, Trausan-Matu, & Stahl, 2005; Stahl, 2006c). The assumption is that the temporality of a chat conversation in terms of references to prior postings and to future ones is similar to life, as "our present is located within a nexus of ties to the past or hopes for the future" (Stahl, 2006c, p. 102). As the experience of the conversation group becomes analogous to a "lived sense of time […] attuned to the larger world outside" (*ibid.*), rhythm also emerges as a determinant for the collaborative experience analysis.

Based on the previous theoretical bridges of polyphony and rhythm between language and music, we have proposed in *7.5* Dialogism and Voice Inter-Animation a series of measures applied on the automatically identified voices, seen as lexical chains combined with semantic relationships. These factors are aimed at modeling the voice synergy effect that induces the rhythm of discourse. Moreover, in *9.2.2* Dialogical Voice Inter-Animation Model we have introduced a collaboration assessment model based on voice inter-animation that highlights the feasibility of using the synergic effect between voices of different participants in order to determine intense collaboration zones.

3.1.3 CSCL Computational Approaches

Instant messenger (chat) has been already used for several years in Computer Supported Collaborative Learning (CSCL) sessions (Stahl, 2006b). Later on subsequent technologies have been adopted in educational scenarios. However, there are very few systems that automatically analyze such conversations and provide feedback to both learners and tutors. The explanation is probably that Natural Language Processing is needed and the existing technologies in computational linguistics are still not mature, especially for analyzing chat conversations, which have many important differences as compared to non-conversational text: 1/ shorter utterances, that can be divided into multiple subsequent interventions; 2/ elliptic formulations; 3/ the frequent use of emoticons and other abbreviations; 4/ intertwining of multiple concurrent discussion threads that decreases the coherence of the discourse seen as adjacent utterances; 5/ the social dimension of the analysis that needs to be taken into consideration.

Several CSCL systems were developed for analyzing interactions in conversations using transcriptions of spoken conversations, logs of instant messenger (chat), forum interventions and even wikis. Within the state of the art we decided to focus on the systems with which we found greater similarities to our wok: *CORDTRA* – activity patterns can be generalized to voice inter-animation patterns, whereas a similar diagram can be used for voice visualization, *DIGALO* – patterns were also used within our developed systems (*A.S.A.P.*, *Ch.A.M.P.* and *PolyCAFe*) for identifying speech acts and inter-animation, *KSV* – social network analysis applied on a 'notes' graph, plus the integration of Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) (see *4.3.2* Semantic Similarity through Tagged LSA) as similarity measure.

CORDTRA (Hmelo-Silver, Chernobilsky, & Masto, 2006) builds diagrams with parallel timelines that enable the user to juxtapose a variety of codes to understand activities (e.g., see Figure 8 that displays discourse, gestural and tool-related codes). Although initially used to evaluate face-to-face collaboration in a problem-based learning (PBL) tutorial, the diagrams were later on integrated in the *eSTEP* system (Hmelo-Silver, Chernobilsky, & Nagarajan, 2005) in order to study online collaborative learning through the evolution of activity patterns – sequential involvement was in the detriment of group effectiveness (Hmelo-Silver et al., 2006).

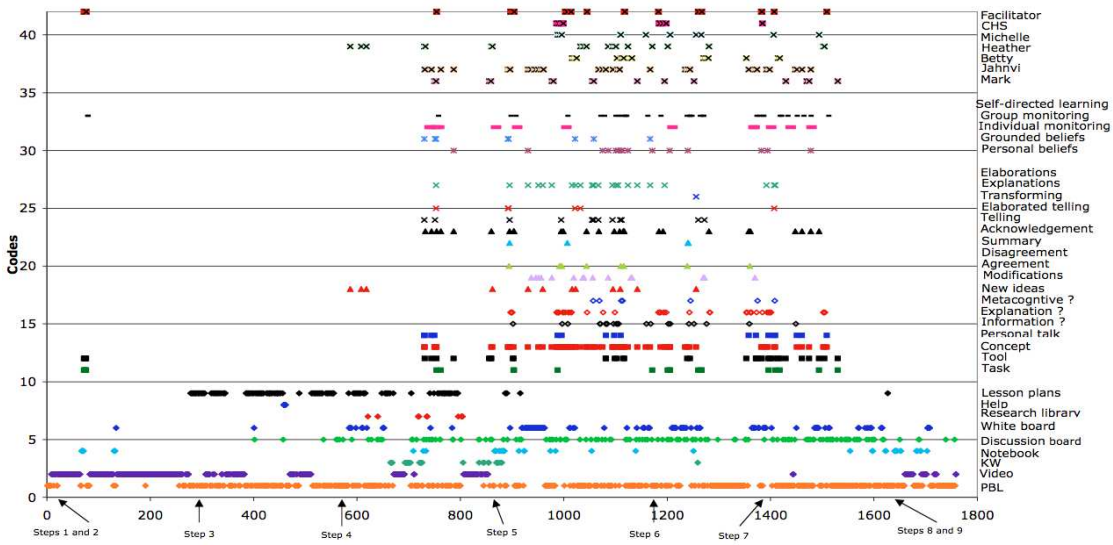


Figure 8. Example of a *CORDTRA* diagram (Hmelo-Silver et al., 2006, p. 1062).

DIGALO and other tools used in the *ARGUNAUT* project (Harrer, Hever, & Ziebarth, 2007) process log files to identify insightful patterns within the learning process through: 1/ explicitly specified patterns and 2/ automatically discovered patterns, matching configurable parameters. Afterwards, pedagogical experts specified patterns of interest by action types, and rules were enforced in order to model different phenomena, ranging from simple, directly observable patterns (e.g., agreement) to quite abstract and complex concepts (e.g., change of opinion, chain of reasoning) (see Figure 9). As drawbacks of this pattern-based approach, 1/ multiple patterns could compete between the same utterances or 2/ patterns and actions might intersperse, generating insignificant scenarios and inhibiting meaningful patterns.

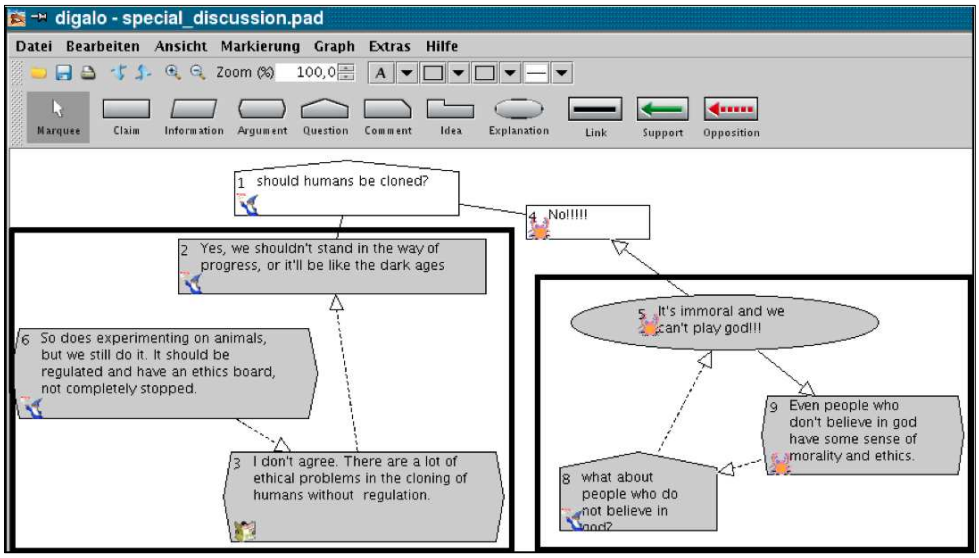


Figure 9. Example of a *DIGALO* discussion map with two identified hits in the chain of reasoning pattern (Harrer et al., 2007, p. 509).

Collaborative Learning

The *Knowledge Space Visualizer* (*KSV*) (Teplovs, 2008) builds a graph of participant 'notes' based on multiple relationships that range from structural (e.g., reply-to, build-on, reference, annotation, contains), authorial, to semantic, based on Latent Semantic Analysis (Landauer & Dumais, 1997) similarity, or derived from researcher coding. A force-directed layout is used for visualizing the generated network (Heer, Card, & Landay, 2005) (see Figure 10). Moreover, *KSV* encourages a cycle of continual analytic improvement, as data is gathered from the *Knowledge Forum* (Scardamalia, 2004) and can be fed back into it. As specificity, *KSV* performs also clustering on nodes and integrate multiple data sources.

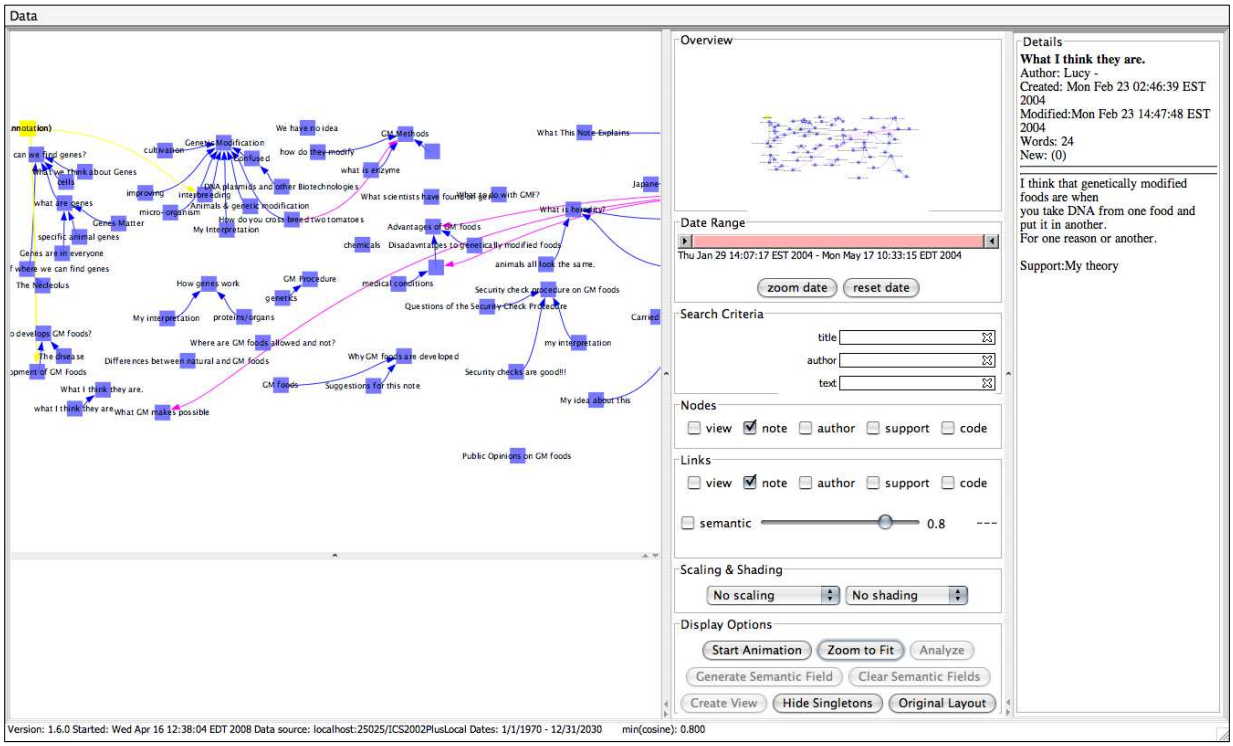


Figure 10. The *Knowledge Space Visualizer* (Teplovs, 2008).

Other remarkable approaches in terms of CSCL, but more distant from ours, include: 1/ *ColAT* (Avouris, Fiotakis, Kahrimanis, Margaritis, & Komis, 2007) that inter-relates multiple resources (e.g., log files, video, audio files, snapshots) for generating interpretative views of activity; 2/ the Scaffold-Argument visualization (Law, Lu, Leng, Yuen, & Lai, 2008), useful for understanding knowledge-building discourse through the use of argument markers and scaffolds; 3/ *COALA* (Dowell & Gladisch, 2007; Dowell, Tscholl, Gladisch, & Asgari-Targhi, 2009), a system based on argumentation schemes used for describing patterns of reasoning in discourse; 4/ *TATIANA* that considers also multimedia utterances (Dyke, Lund, & Girardot, 2009); and 5/ *VMT-Basilica*

(Kumar, Chaudhuri, Howley, & Rosé, 2009), a facility for rapid prototyping CSCL environments, integrating also text classification and conversation agents.

Besides the previous approaches, multiple systems were developed within our research group and their corresponding facilities are presented in detail in the empirical part of the thesis: *Polyphony* (Trausan-Matu, Rebedea, et al., 2007), *A.S.A.P.* (Dascalu, Chioasca, & Trausan-Matu, 2008a) (see *5.1 A.S.A.P.* – Advanced System for Assessing Chat Participants), *Ch.A.M.P.* (Dascalu, Trausan-Matu, & Dessus, 2010b) (see *5.2 Ch.A.M.P.* – Chat Assessment and Modeling Program), *PolyCAFe* (Rebedea et al., 2010; Trausan-Matu, Dessus, et al., 2010; Dascalu, Rebedea, Trausan-Matu, & Armitt, 2011) (see *6 PolyCAFe* – Polyphonic Conversation Analysis and Feedback) and *ReaderBench* (Dascalu, Dessus, et al., in press; Dascalu, Trausan-Matu, et al., in press) (see *9 ReaderBench* (3) – Involvement and Collaboration Assessment through Cohesion and Dialogism). In addition, section *9.4* introduces a detailed comparison between *KSV* and *ReaderBench*.

From a completely different point of view, some of these systems use as underlying discourse structures several kinds of argumentation graphs, some of them use the idea of Toulmin (1958), or more elaborated structures like the contingency graph (Suthers, Dwyer, Medina, & Vatrapu, 2007; Medina & Suthers, 2008), utterance graph (Trausan-Matu, Rebedea, et al., 2007; Dascalu, Rebedea, & Trausan-Matu, 2010; Rebedea, Dascalu, Trausan-Matu, Armitt, et al., 2011) or cohesion graph (Trausan-Matu, Dascalu, & Dessus, 2012; Dascalu, Trausan-Matu, et al., in press).

3.2 Social Network Analysis

A social network is a social structure consisting of entities or individuals that are tied together by one or more types of interdependencies highlighting the relationships between them. The connotations of these relations have different implications in terms of the used perspective: from a sociological point of view ties are built based on an underlying meaning of interaction (e.g., links can reflect friendship, kinship or organizational position relations), whereas from a computational perspective the quantity or the quality of transferred information between actors becomes the evaluation determinants (e.g., number of utterances interchanged in a collaborative environment). In addition, considering also that "patterning of links among the people involved in the development of the field – its social network – is a key to understanding how the field emerged" (Freeman, 2004, p. 9), the intertwining of different social networks analysis (SNA) perspectives, from different domains, that

converged in time is impressive. In the end, the "eclectic hodgepodge made up of anthropologists, geographers, social psychologists, communication scientists, political scientists, historians and mathematicians" (Freeman, 2004, p. 30) integrated with the "clique of sociologists". Nevertheless, despite different viewpoints, social networks denote a change from an individual-centered view to a higher perspective, even a global one, in which the relations between nodes are more relevant than their individual attributes (Pinheiro, 2011).

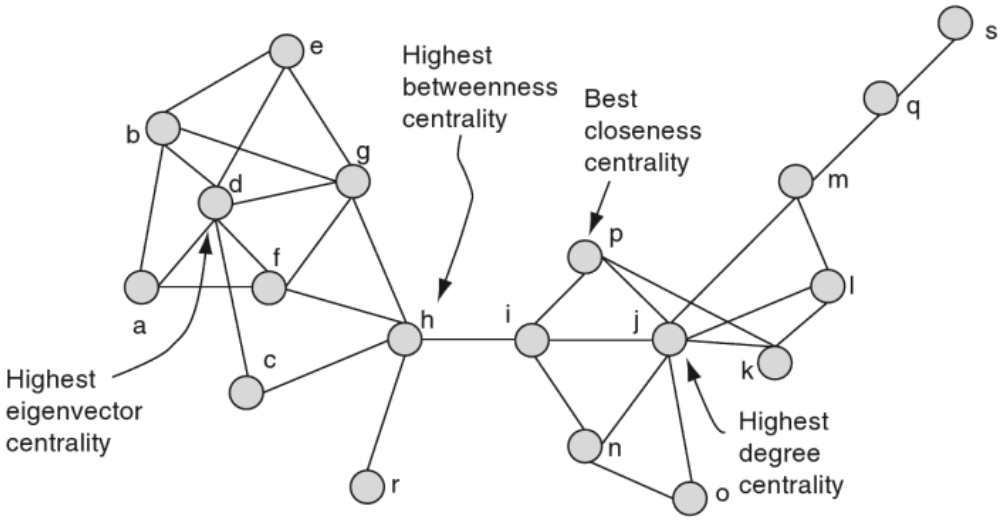


Figure 11. Diverse centrality measures applied on a Social Network Analysis graph example (Ortiz-Arroyo, 2009, p. 30).

From Table 3, the following measures are highlighted in terms of centrality: degree (since we are dealing with an undirected graph, indegree = outdegree), eigenvector centrality, betweenness and closeness centrality

Starting from the previous multi-dimensional approaches, we will focus on the computational perspective of Social Network Analysis (SNA) based on graph theory (Biggs, Lloyd, & Wilson, 1986; Tutte, 2001; Cormen, Leiserson, Rivest, & Stein, 2009), more specifically network theory (Wasserman & Faust, 1994; Newman, 2010). Therefore, in our case, the social network is a graph $G = (V, E)$ that has participants as nodes and the interactions among them as arcs. From this point of view, two different categories of methods can be enforced, each with its own specificities (D'Andrea, Ferri, & Grifoni, 2009): 1/ network data collection focused on the graph's analysis and on determining different evaluation factors (see Table 3 and Figure 11) and 2/ network data visualization and modeling.

Table 3. Main Social Network Analysis factors.

Factor and reference	Formula	Description
Bridge		Closely related to the concept of articulation vertices from graph theory, a bridge represents an individual whose ties provide a unique link between two individuals or clusters.
Structural cohesion (White & Harary, 2001)		Cohesion is measured as the minimal number of nodes that need to be removed from the graph in order to disconnect it (Moody & White, 2003). Closely related to the concept of connectivity from graph theory, it represents from a sociological point of view a formal definition of cohesion in social groups (White & Harary, 2001).
Distance	$d_G(v,t)$	The distance between current node *v* and node *t*, using the minimum length of any path connecting *v* and *t*.
Density	$D = \frac{2\|E\|}{\|V\|(\|V\|-1)}$	Measures the degree to which the current number of edges is close to the maximal number of edges (D = 1 for a complete graph), in other words the sparcity of the graph (Coleman & Moré, 1983).
Indegree		Indegree is defined as the sum of head endpoints in the graph, in-oriented towards a specific node. It is closely linked to the concept of expertise, as more interventions oriented towards a participant usually denote a more knowledgeable individual within the discussion group.
Outdegree		Outdegree is the reflection of indegree as it is composed of the sum of tail endpoints in the graph, out-oriented from a specific node. This is usually a mark of an intense participation within the group, although it might be associated with gregariousness if a high discrepancy exists between the in- and out-degree values.
Closeness Centrality (Sabidussi, 1966)	$C_C(v) = \frac{1}{\sum_{t \in V} d_G(v,t)}$	Closeness reflects centrality in the inverse of the minimal distances between the current node and all the other nodes in the graph. It can be considered a measure of speed in terms of spreading sequentially the information from *v* to all other nodes (Newman, 2005).
Eccentricity or Graph Centrality (Freeman, 1977)	$C_G(v) = \frac{1}{max_{t \in V} d_G(v,t)}$,	Eccentricity measures the relative closeness of a participant to all other individuals, as it considers the maximal distance between node *v* and all the other nodes in the graph.

Factor and reference	Formula	Description
Dangalchev's centrality (Dangalchev, 2006)	$C_C(v) = \sum_{t \in V \setminus v} 2^{-d_G(v,t)}$	Although this centrality measure was firstly introduced into a completely different domain in order to measure network vulnerability as the time required to access a given node while compromising different nodes in the network, this factor is of particular interest as it can be used also for disconnected graphs (as the distance between nodes can be infinity, the inverse exponential function automatically cleans the distances between isolated nodes).
Eigen Centrality (Newman, 2008)	$x_i = \frac{1}{\lambda} \sum_{j \in M(i)} x_j$ $= \frac{1}{\lambda} \sum_{j=1}^{N} A_{i,j} x_j$	Eigen values determine the importance of a participant in a network based on the following principle: a connection to a higher-ranking node is more important than a multiple connections to inferior-ranked nodes. $M(i)$ = set of individuals connected to the i^{th} node; N = total number of nodes; λ – constant.
Shortest Path Betweenness Centrality (Freeman, 1977; Brandes, 2001)	$\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$ $C_B(v)$ $= \sum_{s \neq v \in V} \sum_{t \neq v \in V} \delta_{st}(v)$	Betweenness reflects the number of times a given node acts as a bridge along all shortest paths between pairs of two other nodes. It can be generalized to a measure of control in the communication between two other nodes (Freeman, 1977). $\sigma_{st}(v)$ = number of shortest paths that contain node v; σ_{st} = number of all shortest paths between s and t.
PageRank (L. Page, 2001)	$NR(v)$ $= (1-d) + d$ $\times \sum_{t \in V \setminus v} \frac{NR(t)}{C(t)}$	A node's rank is influenced by the other nodes' ranks that are directly addressing him. Therefore, the messages the node receives and the rank of the nodes transmitting them are the main factors for determining his current ranking. In other words, the more a node is accessed, the more it proves the high value of the information he is transmitting to other nodes, thus his ranking will increase in time. d = 0.85 as an optimal value used for a faster convergence; $C(t)$ = the weight of $t \rightarrow v$.

Besides the computations that can be performed in order to evaluate a social network, visualization plays an important role in understanding and interpreting the obtained results (D'Andrea et al., 2009). Although a multitude of visualizations have been devised in time, the most frequently used and the easiest to understand layouts used for providing the shape of the graph are the following:

- *Force-based layouts* that provide an overview of the social network as a planar representation in which nodes gravitate, having their own mass. The length between nodes is proportional to the strength of the tie and elasticity coefficients are used to

obtain a more realistic model of the network. In the end, in order to obtain an aesthetically pleasant representation, the aim of the visual representation is to minimize edge crossings and the overall network energy (Bannister, Eppstein, Goodrich, & Trott, 2012) (see Figure 12.a). Afterwards multi-level algorithms (graph coursing) can be applied in order to generate a more refined and representative visualizations in which the initial graph is partitioned and nodes are grouped into clusters (see Figure 12.b).

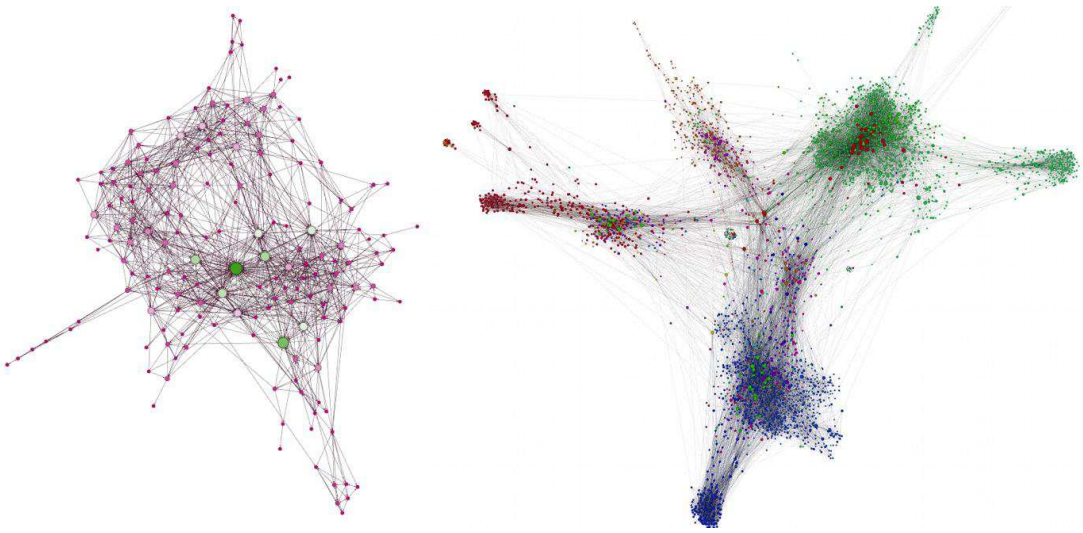


Figure 12. *Gephi* (Bastian, Heymann, & Jacomy, 2009) force-based layouts. a. graph overview; b. coarse graph.

- *Radial layouts* which offer a central or ego-centric perspective (D'Andrea et al., 2009) of an individual within the social network. Therefore, the graph is focused on a central node and his neighbors, whereas concentric levels reflect an increase in distance (see Figure 13).

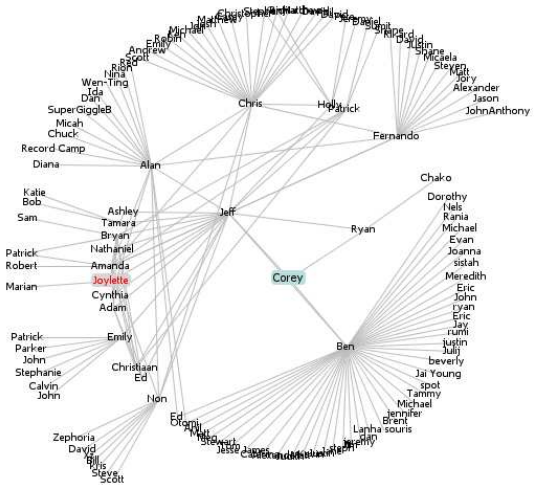


Figure 13. *Prefuse* (Heer et al., 2005) radial layout.

When addressing CSCL, collaboration graphs (Harary, 1979) were introduced for modeling social networks in which vertices represent participants and where two nodes are linked together whenever a collaborative relationship between them exists. As an example, the Erdős collaboration graph (Batagelj & Mrvar, 2000) models a network of mathematicians that are linked if they co-authored a paper (without the limitation of being the only two co-authors, as all combinations of co-author pairs were taken into consideration). More specific to CSCL conversation, nodes are participants in a collaborative environment and ties can be generated based on explicit links between utterances, obtained for example from the explicit referencing facility of the used chat environment (Holmer et al., 2006), and on implicit links, obtained using natural language processing techniques (Trausan-Matu, Rebedea, et al., 2007; Trausan-Matu, Rebedea, & Dascalu, 2010).

Moreover, when modeling interaction between individuals, often captured in their communication networks or more specifically in CSCL environments, SNA can be considered the principal quantitative Learning Analytics (LA) tool (Cooper, 2012) for modeling participant interactions. Therefore, SNA is useful for determining and modeling the importance of a participant in a conversation or within a discussion group, including the visualization of interaction and of centrality in the social network.

3.3 Metacognition and Self-regulation in CSCL

Although self-regulation, guided by metacognition seen as "cognition about cognition" or "knowing about knowing" (Metcalfe & Shimamura, 1994), is specific to individual learning, similar processes are present in collaborative learning as well. The goal of this section is to create relationships between individual and collaborative learning, although at a first glance self-regulated learning (SRL) is focused only on the individual. SRL "emphasizes autonomy and control by the individual who monitors, directs, and regulates actions toward goals of information acquisition, expanding expertise, and self-improvement" (Paris & Paris, 2001, p. 89).

In contrast to the process of self-regulated learning that has been thoroughly analyzed (Vohs & Baumeister, 2011), there is few research addressing the twofold transition towards CSCL: firstly, the shift must be performed towards a computer supported perspective – Computer-based Learning Environments (CBLE) (Chen, 1995), followed by a change from individual to collaborative learning in a technology-enhanced environment. As CBLEs offer important opportunities for fostering

learning (Lajoie & Azevedo, 2006), Winters, Greene, and Costich (2008, p. 440) argue that "different learner and task characteristics (e.g., prior knowledge, goal orientation, learner control) and types of learner support are related to students' SRL when using CBLEs". Therefore technology can be considered a mediator, although its efficacy reflected in particular SRL processes needs to be further studied (Winters et al., 2008). In addition, from a science educational perspective, Kali and Linn (2007) emphasize eight pragmatic design principles from the Design Principles Database that most likely support learning through applying technology features: 1/ communicate the diversity of science inquiry, 2/ connect to personally relevant examples, 3/ provide students with templates to organize ideas, 4/ provide knowledge representation tools, 5/ enable three-dimensional manipulation, 6/ *encourage learners to learn from others* – emphasis on collaboration by using the *eStep* system (Derry et al., 2005) in which learners are presented with a classroom dilemma that needs to be collaboratively solved, 7/ enable manipulation of factors in models and simulations and 8/ encourage reflection.

While performing the transition from technology-enhanced learning (TEL) to CSCL, Glahn, Specht, and Koper (2009) found that 1/ a tag cloud visualization of a learner's 'freeform tags' can stimulate reflective meta-cognitive processes and 2/ new tools that support self-directed learners may take advantage of the concepts of situated learning (Lave & Wenger, 1991) focused on the social dimension of learning – e.g., the use of *ReScope* tag cloud (Glahn, Specht, & Koper, 2008) based on the concepts and process dimensions. With emphasis on chats accompanied by whiteboard tools, Pata and Sarapuu (2003) studied meta-communicative scaffolding patterns required as support in performing expressive modeling tasks. Their conclusions were encouraging in terms of the suggested scaffolding tools required by learners in order to develop effective regulation strategies: "i) directing students towards using chat room for explaining the topics related to content generation, task regulation and interaction planning, while referring to the actions carried out on the whiteboard area; and ii) fostering self-explaining activities that will reflect thinking processes to the team-members" (Pata & Sarapuu, 2003, p. 1129).

Järvenoja and Järvelä (2009) have demonstrated that individual group members from a collaborative learning environment that socially share learning tasks can play a leading role in activating motivation regulation. Motivation is one of the principal elements learners need to plan, monitor, regulate and control, besides cognition, behavior and context (Boekarts, Pintrich, & Zeidner, 2000). Moreover, collaborative knowledge construction and joint metacognitive regulation may also stimulate new

strategies for motivation regulation (Hurme, Palonen, & Järvelä, 2006). In addition, "successful engagement in collaborative learning presumes norms that allow members to feel safe, take risks and share ideas. This actually involves core processes of self-regulated learning; effective use of learning strategies to participate in collaborative interactions; metacognitive control and regulation of motivation and emotions" (Järvelä, Hurme, & Järvenoja, 2011, p. 20). As conclusion, the broadening of the CSCL perspective to self-regulated learning has beneficial effects in terms of the learner that could become more actively involved and motivated, improving nevertheless his/her skills in social learning practices (Järvelä et al., 2011).

4 Computational Discourse Analysis

As previous chapters were overall oriented towards comprehension and productions from the perspectives of individual and collaborative learning, this chapter is focused on presenting automatic discourse analysis models and natural language processing techniques that ground a computational and quantifiable perspective of cohesion and coherence and that greatly impact the underlying functionalities of our developed systems (*A.S.A.P.*, *Ch.A.M.P.*, *PolyCAFe* and *ReaderBench*).

4.1 Measures of Cohesion and Local Coherence

From a computational viewpoint, we limit the perspective of coherence and cohesion (see *2.1.1* Coherence and Cohesion) to *lexical and semantic cohesion* and *local coherence* that captures text organization at the level of sentence to sentence transitions, further necessary to achieve global coherence (Lapata & Barzilay, 2005). In this computational context, *cohesion* is reflected in the linguistic form of discourse (McNamara et al., 2010) and is often regarded as an indicator of its structure. More specifically, cohesion can derive from: 1/ discourse connectedness through cue words or phrases (e.g., "but", "because") as relations between sentences (e.g., explanation, contrast); 2/ referencing expressions that reflect the status of an entity in the discourse and can be identified through co-reference resolution (Jurafsky & Martin, 2009; Raghunathan et al., 2010); 3/ lexically or semantically related words obtained from semantic distances in ontologies (Budanitsky & Hirst, 2006) (see *4.3.1* Semantic Distances and Lexical Chains), cosine similarity in vector spaces from Latent Semantic Analysis (Landauer, Foltz, et al., 1998) (see *4.3.2* Semantic Similarity through Tagged LSA) or through topic relatedness in Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) (see *4.3.3* Topic Relatedness through Latent Dirichlet Allocation). Aligned with the previous definition are also the two measures of textual cohesion proposed by Graesser et al. (2004), frequently used in automated discourse analysis: *referential cohesion* (the degree to which words, concepts or

phrases are related or repeated across the text) and *causal cohesion* (marked by the explicit use of connectives – e.g., "since", "because", "therefore", "the cause of" or "as a consequence").

Coherence, on the other hand, is much more difficult to express from a computational perspective as multiple levels that simultaneously relate discourse elements need to be taken into consideration (Grosz & Sidner, 1986). Moore and Pollack (1992) focus on two levels in particular: 1/ the *informational level*, mostly centered on causal relations between utterances, weakly related to the linguistic form and difficult to model in comparison to previous links between words; and 2/ the *intentional level*, aimed at the changes in the discourse participants' mental states, superficially visible in the linguistic form and extremely difficult to model in terms of computational analysis. Moreover, the same study highlights also a problem of the rhetorical structure theory (Mann & Thompson, 1987a) (see *4.2* Discourse Analysis and the Polyphonic Model) that is limited to a single, preferred rhetorical relation between consecutive discourse elements, whereas coherence should be modeled as an overlap of multiple relations between the same text spans, but at different levels. Nevertheless, while addressing the informational level, coherence is most frequently accounted by: lexical chains (Morris & Hirst, 1991; Barzilay & Elhadad, 1997; Lapata & Barzilay, 2005) (see *4.3.1* Semantic Distances and Lexical Chains), centering theory (Grosz et al., 1995; Miltsakaki & Kukich, 2000) (see *4.2* Discourse Analysis and the Polyphonic Model) in which coherence is established via center continuation, or Latent Semantic Analysis (Foltz, Kintsch, & Landauer, 1993, 1998) (see *4.3.2* Semantic Similarity through Tagged LSA) used for measuring the cosine similarity between adjacent phrases; in the end, overall coherence is considered the mean value of the previous semantic similarities. Nevertheless, from a computational perspective and through its intrinsic nature consisting of a bag-of-words approach, LSA fundamentally supports cohesion and not coherence.

4.2 Discourse Analysis and the Polyphonic Model

Discourse may be defined as "a coherent structured group of sentences" (Jurafsky & Martin, 2009, ch. 21) that in NLP is usually considered different in monologues and dialogues. However, in both cases the same idea of an emitter–receiver channel is used, the difference being the uni- respectively bi-directional communications (Trausan-Matu & Rebedea, 2010). Therefore, one-way, speaker-listener directed models of communication are considered in monologues (Jurafsky & Martin, 2009). The usual way of analyzing discourse in this case is the segmentation of text, the search for different relationships among segments, the measurement of coherence and obtaining some discourse

abstractions like co-references or summaries. In this context, cohesion seen as lexical, grammatical and semantic links between textual fragments becomes a central element, whereas coherence is considered as granted, in different degrees, when analyzing texts. On the other hand, the detection of local relations can be used for measuring coherence. Some structures are searched, as the Rhetorical Structure Theory (RST) (Mann & Thompson, 1987b, 1988), which considers a hierarchical decomposition of a text. Centering Theory (Grosz et al., 1995) and co-reference resolution systems (Jurafsky & Martin, 2009) may be also considered (Trausan-Matu & Rebedea, 2010).

On the other hand, dialogue analysis has as prototype phone-like (or face-to-face) conversations. A typical approach starts from analyzing local, two-participant data and tries to identify speech acts, dialog acts and afterwards, adjacency pairs (Jurafsky & Martin, 2009). Even if there are attempts to analyze conversations with multiple participants, considering a more global, collaboration-based perspective, like transacts (Joshi & Rosé, 2007), the approach is also based on a two interlocutors' model (Trausan-Matu & Rebedea, 2010).

In terms of discourse analysis, probably the most known discourse theories belong to Hobbs (1985), Grosz et al. (1995) or Mann and Thompson (1987b). Hobbs' theory is based on considering semantic coherence relations – "a set of binary relations between a current utterance and the preceding discourse" (Hobbs, 1978, p. 2) – and on using abduction inferences in formal logic (Hobbs, 1979, 1985). "Coherence thus plays a role beyond sentence boundaries analogous to the role played by grammaticality within sentences. It is the mortar with which extended discourse is constructed." (Hobbs, 1979, p. 69). Also, of particular interest is the phenomenon of topic drifting observed in spoken conversations – although adjacent segments are coherent, the end of the conversation is significantly different from the starting point – that is mainly induced by three mechanisms: sematic parallelism, chained explanations and metatalk (Hobbs, 1990).

Rhetorical Structure Theory (RST) (Mann & Thompson, 1987b) identifies hierarchical rhetorical structures between text spans (defined as any contiguous interval of text), classified as nuclei or satellites in accordance to their importance, that is built using a limited set of rhetorical schemas (patterns) like antithesis and concession, elaboration, enablement and motivation, interpretation and evaluation, restatement and summary, etc. The theory requires the fulfillment of 4 constraints for a successful RST analysis (Mann & Thompson, 1987b): 1/ completeness as coverage of the entire text,

2/ connectedness focusing on the recursive division of text spans, 3/ uniqueness as each relation is applied on different text spans and 4/ adjacency as adjoined text spans are consecutive.

From a different perspective, coherence is obtained in the centering theory (Grosz et al., 1995) at both local (coherence among the utterances in a given segment) and global levels (coherence with other segments of the discourse), centered on two different aspects: intentional and attentional states, which together with the linguistic structure of an utterance sequence, form a tripartite organization. An intentional structure should be present in each discourse, assuring that discourse is rational. This structure is built from intentions (purposes) and, sometimes, from the beliefs of the author of the discourse (or of each participant in a conversation) and from relationships among linguistic segments (Grosz et al., 1995). In addition, two types of centers are identified: backward-looking and forward-looking. Continuation of the discourse is modeled through an ordered set of forward-looking centers defined at utterance level plus a single back-looking center (except for the first utterance of the discourse segment), "that provides a coherent link to the previous utterances by being coreferential with one of the forward-looking centers of that utterance" (Gordon, Grosz, & Gillom, 1993, p. 312).

On the other hand, the polyphonic theory (Trausan-Matu, Stahl, & Zemel, 2005; Trausan-Matu & Stahl, 2007; Trausan-Matu & Rebedea, 2009; Trausan-Matu, 2010c; Trausan-Matu, Rebedea, et al., 2010) follows the ideas of Koschmann (1999) and Wegerif (2005) and investigates how Bakhtin's theory of polyphony and inter-animation (Bakhtin, 1981, 1984) (see *3.1.2* Bakhtin's Dialogism) can be used for analyzing the discourse in chat conversations with multiple participants. In phone and face-to-face dialogs only one person usually speaks at a given moment in time, generating a single thread of discussion. This is, of course, determined by the physical, acoustical constraints (if two or more persons are speaking in the same moment, it is impossible to understand something). In chat environments, like the one used in the Virtual Math Teams (VMT) project (Stahl, 2009a), any number of participants may write utterances at the same time. As discussed in a previous section, the VMT environment offers also explicit referencing facilities that allow the users to indicate to what previous utterance(s) they refer to (see *3.1.1* Chats as Support for Social Cognition). This facility is extremely important in chat conversations with more than two participants because it allows the existence of several discussion threads in parallel. Moreover, the co-occurrence of several threads gives birth to inter-animation, a phenomenon similar to polyphony, where several voices jointly play a coherent piece as a whole (Trausan-Matu, Rebedea, et al., 2007; Trausan-Matu & Rebedea, 2009).

Bakhtin (1984) emphasized that polyphony occurs in any text. He considered that dialog characterizes any text, that "our speech, that is, all our utterances (including creative works), is filled with others' words" (Bakhtin, 1986, p. 89). The voice becomes a central concept, has a more complex meaning. A voice is not limited to the acoustic dimension, it may be considered as a particular position, which may be taken by one or more persons when emitting an utterance, which may have both *explicit*, similar to those provided by the VMT chat environment (Stahl, 2009a), and *implicit* links (for example, lexical chains, co-references or argumentation links) and influence other voices. Each utterance is filled with 'overtones' of other utterances (Stahl, 2009a). Moreover, by the simple fact that they co-occur, voices are permanently inter-animating, entering in competition, generating multi-vocality in any conversation and even in any text (in Bakhtin's dialogic theory everything is a dialog) or, as Bakhtin calls it, a "heteroglossia, which grows as long as language is alive" (Bakhtin, 1981, p. 272).

The ideas of Bakhtin drive to a musical metaphor for discourse and for learning: "the voices of others become woven into what we say, write, and think" (Koschmann, 1999, p. 308). Therefore, for analyzing discourse in chats the aim shifts towards investigating how voices are woven, how themes and voices inter-animate in a polyphonic way (Trausan-Matu, Stahl, et al., 2007). This is important not only for understanding how meaning is created but also for trying to design tools for support and evaluation. Figure 14 presents the inter-animation of voices within a chat conversation and their evolution in time, following a pattern first described by Trausan-Matu et al. (2005); the longest two voices are represented by the linked curly lines. As it can be observed, several threads can co-appear in parallel and even the same participant may participate to more than one discussion thread within a given timeframe (e.g. John, at utterance 19, approves and elaborates Tim's intervention, while in the following utterance represents an approval of Adrian's utterance 18) (Trausan-Matu, 2010c). Therefore, this co-presence of multiple discussion threads and their inter-influences models voice inter-animation towards achieving polyphony.

The polyphonic model focuses on the idea of identifying voices in the analysis of discourse and building an internal graph-based representation, whether we are focusing on the utterance graph (Trausan-Matu, Rebedea, et al., 2007) or the cohesion graph (Trausan-Matu, Dascalu, & Dessus, 2012; Dascalu, Dessus, et al., in press) (see *7.2* Cohesion-based Discourse Analysis). For this aim, links between utterances are analyzed using adjacency pairs, repetitions, lexical chains, speech and argumentation acts or cohesive links, a graph is built from which discussion threads are identified.

Nevertheless, in both internal representations, lexical or semantic cohesion between any two utterances seen as explicit communicative acts can be considered the central liaison between the analysis elements within the graph. Cohesion can be expressed as the "distance" between the utterance boundaries (Dong, 2005) and can be computed by various means of semantic similarity, including semantic distances in ontologies (see *4.3.1* Semantic Distances and Lexical Chains), latent vector space representations (see *4.3.2* Semantic Similarity through Tagged LSA) or topic models (see *4.3.3* Topic Relatedness through Latent Dirichlet Allocation).

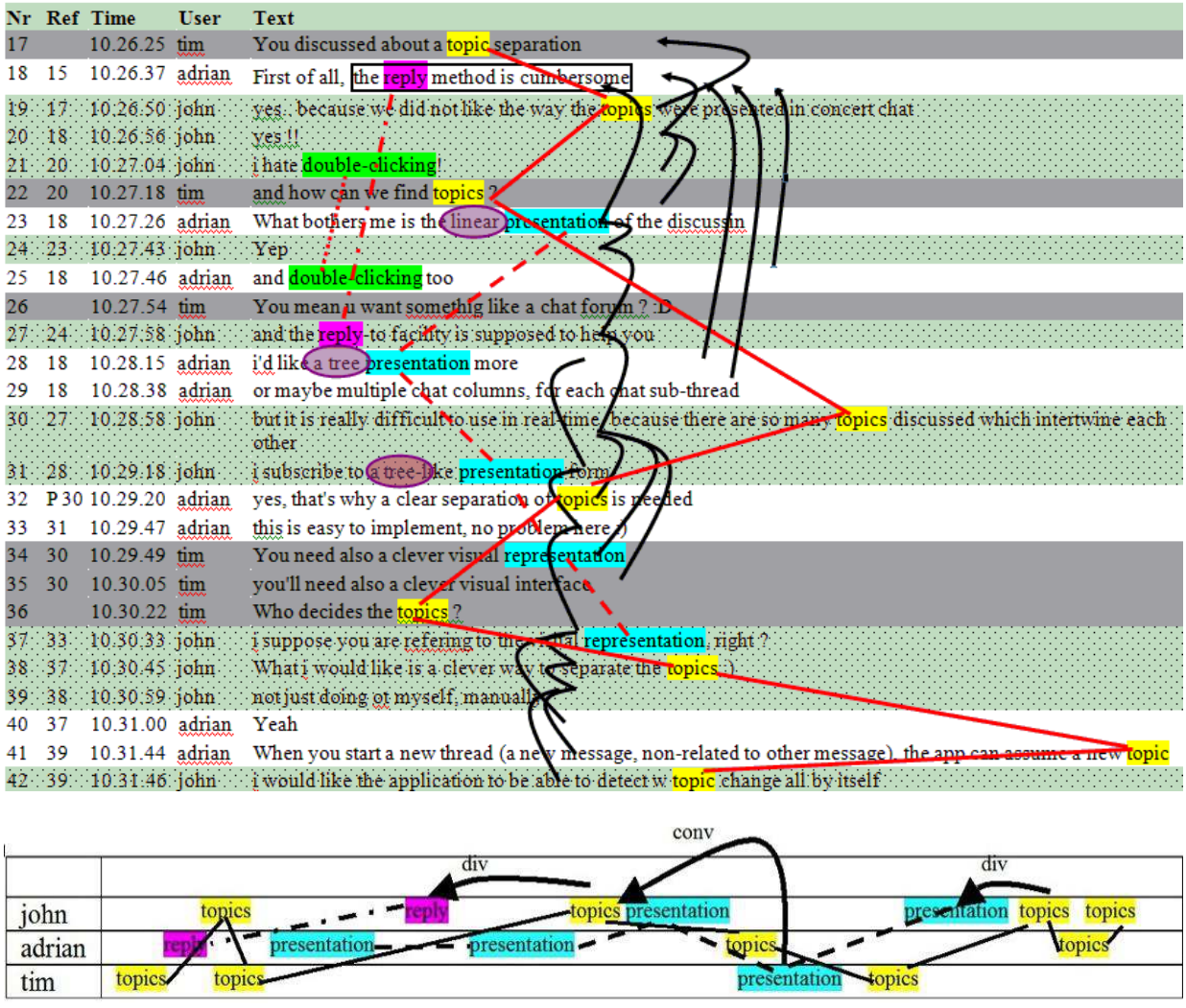


Figure 14. Inter-animation of voices within a chat (Trausan-Matu, Stahl, et al., 2007, pp. 69-70).

As the initial polyphonic model used the utterance graph (Trausan-Matu, Rebedea, et al., 2007) and the cohesion graph (Trausan-Matu, Dascalu, & Dessus, 2012; Dascalu, Dessus, et al., in press), which can be seen as a generalization, is presented in detail in *7.2* Cohesion-based Discourse Analysis, we will focus on providing a comprehensive view of the polyphonic model, using as underlying representation the utterance graph. This internal structure is built upon two types of links between

utterances: explicit and implicit. Participants add manually explicit links during their chat sessions by using a facility from the conversation environment – e.g., Concert Chat (Holmer et al., 2006). On the other hand, implicit links are automatically identified by means of co-references, repetitions, lexical chains and inter-animation patterns (Trausan-Matu et al., 2005; Trausan-Matu & Rebedea, 2010). In the resulted graph, each utterance is a node and the weights of edges are given by the similarity between the utterances. The orientation of each edge follows the timeline of the chat and the evolution of the discussion in time. Starting from the previous graph, a thread can be easily identified as a logical succession of explicitly or implicitly inter-linked utterances. Moreover, the primary extension of each utterance is its inner voice that inter-twines with other voices from the same thread or from different ones, but with less strength. A new intervention or a new utterance in terms of units of analysis can be clearly expressed as a voice and aspects that need to be addressed include: degree of interconnection in terms of cohesion with other utterances, relevance within the discourse or future impact in the overall discussion.

Starting from Bakhtin (1984) perspective of discourse analysis, each identified voice may become more or less powerful than the others and may influence the others. Among chat voices there are sequential and transversal relations, highlighting a specific point of view in a counterpointal way, as mentioned in previous work (Trausan-Matu & Rebedea, 2009). The co-occurrence of several voices which enter in dialogue is a phenomenon considered by Bakhtin to be universal, present in any text, not only in conversations: "Life by its very nature is dialogic … when dialogue ends, everything ends" (Bakhtin, 1984, p. 294). Bakhtin moves the focus of analysis from sentences to utterances in an extended way, in which even an essay contains utterances and is, at its turn, an utterance. Moreover, each utterance is filled with 'overtones' that contain the echoes and influence of other previous utterances.

A voice is generated by an utterance with effects (echoes) on the subsequent utterances via explicit and implicit links. Moreover, by the simple fact that they co-occur, voices are permanently interacting, overlapping and inter-animating, entering in competition, and generating multivocality in any conversation. The ideal situation of a successful conversation or a coherent discourse is achieved when the voices are entering inter-animation patterns based on the discussion threads they are part of (Trausan-Matu et al., 2005).

Moreover, of particular interest is the multi-dimensionality of the polyphonic model (Trausan-Matu, 2013). Firstly, the *longitudinal* dimension is reflected in the explicit or implicit references between utterances, following the conversation timeline. This grants an overall image of the degree of inter-animation of voices spanning the discourse, which can later on be particularized as collaboration, seen as the interactions between multiple participants of the conversation reflected in their voices. Secondly, *threading* highlights voices evolution in terms of the interaction with other discussion threads. Thirdly, the *transversal* dimension is useful for observing a differential positioning of participants, when a shift of their point of interest occurs towards discussing other topics. In the end, this combination of continuity (longitudinal dimension) versus juxtaposition (transversal dimension) of voices, respectively centrifugal versus centripetal forces exerted by participants in terms of covered concepts generates polyphony.

In addition, the co-presence of multiple voices in the same time inherently generates consonances and dissonances, similarly to the polyphonic musical case. In this context, these inter-animation effects of consonance and dissonance in voices overlap can be perceived as centripetal and centrifugal forces tightly correlated in the trend of achieving discourse coherence. The weaving of the voices all along the longitudinal time dimension and meanwhile their consonance/dissonance on the transversal dimension is similar to the case of polyphonic music (Trausan-Matu et al., 2006): "The deconstructivist attack […] – according to which only the difference between difference and unity […] can act as the basis of a differential theory […] – is the methodical point of departure for the distinction between polyphony and non-polyphony." (Mahnkopf, 2002, p. 39)

From a computational perspective, until recently, the goals of discourse analysis in existing approaches oriented towards conversations analysis were to detect topics and links (Adams & Martell, 2008), dialog acts (Kontostathis et al., 2009), lexical chains (Dong, 2006) or other complex relations (Rosé et al., 2008) (see *3.1.3* CSCL Computational Approaches). The polyphonic model takes full advantage of term frequency – inverse document frequency *Tf-Idf* (Adams & Martell, 2008; A. P. Schmidt & Stone), Latent Semantic Analysis (Dong, 2006; A. P. Schmidt & Stone), Social Network Analysis (Dong, 2006), Machine Learning (e.g., Naïve Bayes (Kontostathis et al., 2009), Support Vector Machines and Collin's perceptron (Joshi & Rosé, 2007), the *TagHelper* environment (Rosé et al., 2008) and the semantic distances from the lexicalized ontology *WordNet* (Dong, 2006; Adams & Martell, 2008). The model starts from identifying words and patterns in utterances that

are indicators of cohesion among them and, afterwards, performs an analysis based on the graph, similar in some extent to a social network, and on threads and their interactions.

As conclusion, the polyphonic discourse analysis model, built on Bakhtin's dialogism and supported by multiple natural language processing techniques (presented in detail in *4.3* Natural Language Processing Techniques) can be considered a viable representation of discourse, with emphasis on the analysis of multi-participant conversations for which classic approaches are not appropriate. Moreover, initial validations performed by Trausan-Matu (2011) showed that the results of the polyphonic analysis were close to those of tutors, whereas its extension in terms of assessing collaboration (see *9.2* Collaboration Assessment) proves its applicability.

4.3 Natural Language Processing Techniques

While addressing natural language processing techniques (Manning & Schütze, 1999), of particular interest is to what extent computational models of semantic memory (Cree & Armstrong, 2012) grasp underlying semantic relations and meanings of concepts from texts, and how these models can be effectively used to measure cohesion between textual fragments (Bestgen, 2012). In this context, three complementary approaches are most remarkable: 1/ semantic distances in ontologies (Budanitsky & Hirst, 2006), 2/ semantic vector spaces extracted through Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) and 3/ probabilistic topics modeling by using Latent Dirichlet Allocation (Blei et al., 2003), presented in detail in the current section and integrated in various developed systems. The presentation of each approach offers a broad perspective of the method and of the used resources, particularities, possible improvements and drawbacks.

4.3.1 Semantic Distances and Lexical Chains

As knowledge can be formally represented as a conceptualization consisting of objects, concepts or other entities presumably related to an area of interest and of relationships linking them together (Genesereth & Nilsson, 1987), an ontology can be seen as an "explicit specification of a conceptualization" (Gruber, 1993). Therefore, an ontology consists of a set of concepts specific to a domain and of the relations between pairs of concepts. Starting from the representation of a domain, we can define various distance metrics between concepts based on the defined relationships among them and later on extract lexical chains, specific to a given text that consist of related/cohesive concepts spanning throughout a text fragment or the entire document.

A Lexicalized Ontologies and Semantic Distances

One of the most commonly used resources for English sense relations in terms of lexicalized ontologies is the *WordNet* lexical database (Miller, 1995; Fellbaum, 1998; Miller, 2010) that consists of three separate databases, one for nouns, a different one for verbs, and a third one for adjectives and adverbs. *WordNet* groups words into sets of cognitively related words (synsets), thus describing a network of meaningfully inter-linked words and concepts. Therefore, synonymy is the main relation between words that are now grouped into unordered sets that also include a brief description or gloss, useful for word sense disambiguation (WSD) (Navigli, 2009).

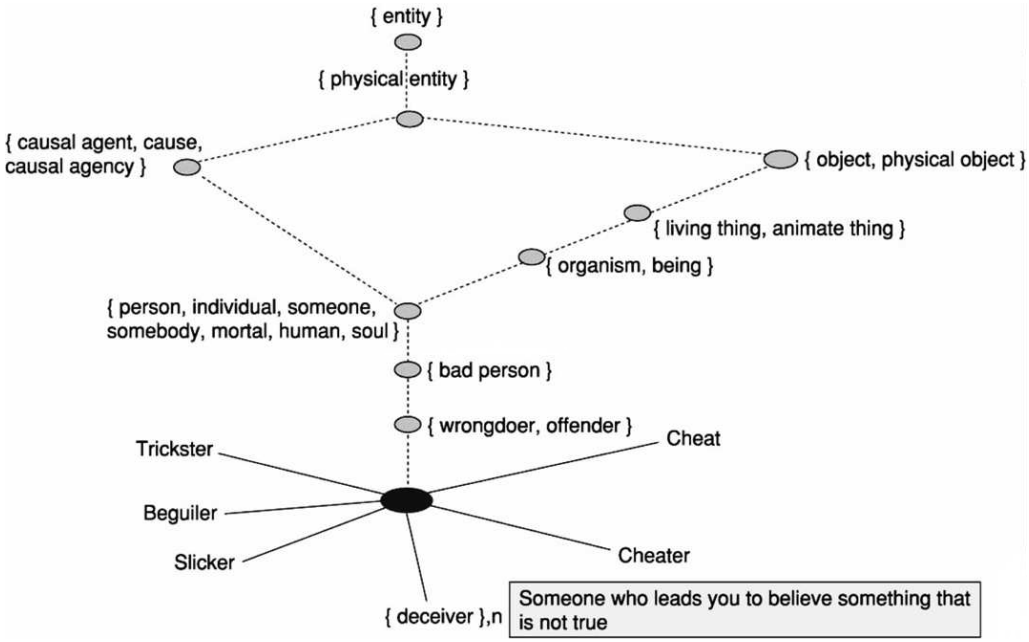


Figure 15. *WordNet* noun tree reflecting semantic/hierarchical relations (Fellbaum, 2005, p. 666).

In addition, *WordNet* is built using the principle of "cognitive plausibility" as the organization of words mimics cognitively related concepts (Miller, 1998; Emond, 2006). This principle of plausibility is based on three hypotheses: *separability* – "lexical knowledge is independent from other language related knowledge"; *patterning* – "relations and patterns between lexical entities are central to natural language processing" and *comprehensiveness* – "any computation model of human language processing should have a store of lexical knowledge as extensive as people do" (Miller, 1998; Emond, 2006).

Synsets are interconnected using semantic relations that vary based on the underlying part-of-speech (see Figure 15 and Table 4). In addition, the internal organization of nouns and verbs uses a hierarchy built on "IS A" relationships and the links between synsets can be regarded as specialization relations

between conceptual categories, aligning the perspectives of *WordNet:* lexical database versus lexicalized ontology. As an overview of *WordNet*, each database consists of a set of lemmas annotated with a set of corresponding senses, covering in the 3.0 version approximately 117k nouns, 11k verbs, 22k adjectives and 5k adverbs; the average noun has 1.23 senses, while verbs have 2.16 senses on average.

Table 4. Word part-of-speech and relations between synsets in *WordNet* (Fellbaum, 2005).

Word part-of-speech	Available relations between synsets
Noun	*hypernymy* – "is a" generalization
	hyponymy – "is a" specialization
	coordination/sibling – concepts share a hypernym
	holonymy – "is a part of" generalization
	meronymy – "is a part of" specialization
Verb	*Entailment relationships*
	troponymy- one activity expresses a particular manner of the other
	backward entailment, presupposition and *cause*
Adjective	*Descriptive adjective*
	direct antonymy and *indirect antonymy*
	Relational adjective
	related noun
Adverb	*base adjective*

Regarding other freely available similar resources, *WordNet Libre du Francais – WOLF* (Sagot, 2008; Sagot & Darja, 2008) is the best French alternative that uses the XML file format developed within the IST-2000-29388 BalkaNet – Design and Development of a Multilingual Balkan WordNet project (http://www.dblab.upatras.gr/balkanet/).

Besides word sense disambiguation, *WordNet* or similar resources are useful for determining the relatedness between concepts through semantic distances (Budanitsky & Hirst, 2001; Pedersen, Patwardhan, & Michelizzi, 2004; Budanitsky & Hirst, 2006; Wang & Hirst, 2011) (see Table 5), query expansion using lexical-semantic relations (Voorhees, 1994; Moldovan & Mihalcea, 2000; Navigli & Velardi, 2003) or the identification of speech acts (Yeh, Wu, & Chen, 2008; Trausan-Matu & Rebedea, 2010). Although multiple semantic distances exist and more can be added to the list presented in Table 5, there is no clear measure that best fits all analysis scenarios as "lexical

semantic relatedness is sometimes *constructed* in context and cannot always be determined purely from an a priori lexical resource such as *WordNet*" (Murphy, 2003; Budanitsky & Hirst, 2006).

Nevertheless, we must also present the limitations of *WordNet* and of semantic distances, with impact on the development of subsequent systems (see *6 PolyCAFe* – Polyphonic Conversation Analysis and Feedback and *7 ReaderBench* (1) – Cohesion-based Discourse Analysis and Dialogism): 1/ the focus only on common words, without covering any special domain vocabularies; 2/ reduced extensibility as the serialized model makes difficult the addition of new domain-specific concepts or relationships; 3/ most relations are between words with the same corresponding part-of-speech, significantly reducing the horizon for comparing the semantic relatedness between concepts; 4/ semantic problems or limitations, specific to a given context, that require additional cleaning – the *OntoClean* approach (Oltramari, Gangemi, Guarino, & Masolo, 2002) and 5/ the encoded word senses are too fine-grained even for humans to distinguish different valences of particular concept senses, reducing the performance of WSD systems. For the later granularity issue, multiple clustering methods that automatically group together similar senses of the same word have been proposed (Agirre & Lopez, 2003; Navigli, 2006; Snow, Prakash, Jurafsky, & Ng, 2007). In addition, when considering *WOLF* in which glosses are only partially translated, integrating in the end a mixture of both French and English definitions, only a limited number of semantic distances are applicable (e.g., path length, Leacock-Chodorow's normalized path length or Wu-Palmer as the most representative).

Table 5. Semantic distances applied on *WordNet.*

Name and reference	Formula	Description
Path length	$l(c_1, c_2)$	The length of the shortest path between two concepts/synsets.
Depth	$d(c_1) = l(c_1, root)$	The length of the path from the current concept to the global root.
Hirst-St-Onge (Hirst & St-Onge, 1997)	$rel_{HS}(c_1, c_2) = C - l(c_1, c_2) - k \times dir$	Two words are considered semantically related if the path is not too long and its direction does not change too often (*dir* – number of direction changes; *k*, *C* – constants).
Leacock-Chodorow (Leacock & Chodorow, 1998)	$sim_{LC}(c_1, c_2) = -log\frac{l(c_1, c_2)}{2D}$	The path length is normalized by the overall depth *D* of the ontology.
Resnik (Resnik, 1995)	$sim_R(c_1, c_2) = -log(p(lso(c_1, c_2))$	Similarity is expressed as the information content of their lowest super-ordinate ($lso(c_1,c_2)$ – most specific common sub-summer; $p(c)$ – probability of occurrence of synset *c* in a specific corpus).
Jiang-Conrath (Jiang & Conrath, 1997)	$dist_{JC}(c_1, c_2) = 2 \times log\left(p(lso(c_1, c_2))\right) - log(p(c_1)) - log(p(c_2))$	Besides the consideration of the most specific sub-summer, the information content of the two nodes also plays an important role in estimating the inverse of similarity.
Lin (Lin, 1998)	$sim_L(c_1, c_2) = \frac{2 \times log\left(p(lso(c_1, c_2))\right)}{\log(p(c_1)) + log(p(c_2))}$	The measure follows the idea of similarity between objects, combined with $dist_{JC}$.
Wu-Palmer (Z. Wu & Palmer, 1994)	$sim_{WP}(c_1, c_2) = \frac{2 \times d(lso(c_1, c_2))}{l(c_1, lso(c_1, c_2)) + l(c_2, lso(c_1, c_2)) + 2 \times d(lso(c_1, c_2))}$	Conceptual similarity is a scaled metric perceived in comparison to a global depth.
Lesk (Banerjee & Pedersen, 2002)	$sim_{Lesk}(c_1, c_2)$	Similarity is determined as an adaptation of the Lesk (1986) approach to WordNet by using the overlap between concept descriptions or glosses.

B Building the Disambiguation Graph

Lexical chaining derives from textual cohesion (Halliday & Hasan, 1976) and involves the selection of related lexical items in a given text (e.g., starting from Figure 15, the following lexical chain could be generated if all words occur in the initial text fragment: "cheater, person, cause, cheat, deceiver, …"). In other words, the lexical cohesive structure of a text can be represented as lexical chaining that consists of sequences of words tied together by semantic relationships and that can span across the entire text or a subsection of it. The identified lexical chains are independent of the grammatical structure of the initial text and, in effect, the contained concepts from each chain capture a portion of the cohesive structure of the text. A lexical chain can provide a context for the resolution of an ambiguous term and enable identification of the concept that the term represents. In a particular manner, the lexical cohesive relationships between words can be established using a lexicalized ontology – *WordNet* (Miller, 1995; Fellbaum, 1998) or *WordNet Libre du Francais – WOLF* (Sagot, 2008). Since first proposed (Morris & Hirst, 1991), lexical chains have been used in a variety of applications in the fields of Information Retrieval (IR) (Manning, Raghavan, & Schütze, 2008) and Natural Language Processing (Manning & Schütze, 1999), most notably for word disambiguation (Galley & McKeown, 2003), detection of malapropisms (Hirst & St-Onge, 1997) and text summarization (Barzilay & Elhadad, 1997; Silber & McCoy, 2003).

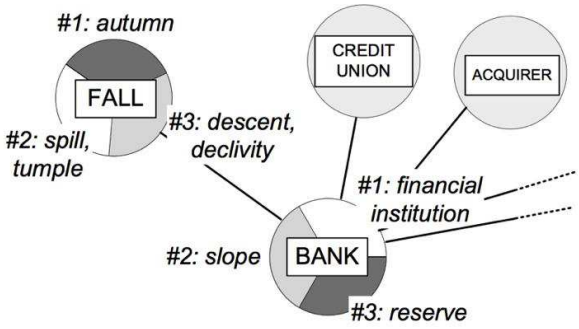


Figure 16. Disambiguation graph example (Galley & McKeown, 2003, p. 1487).
Highlighting a possible implicit representation of word-sense combinations (#*n* denotes a word-sense) with all edge weights equal to 1

Once the pre-processing of a given text is completed (splitting, tokenizing, part of speech tagging, parsing, named entity recognition, co-reference resolution) (Manning & Schütze, 1999), the disambiguation graph (see Figure 16) can be built in linear time (Galley & McKeown, 2003). In this kind of graph, nodes represent word instances and weighted edges represent semantic relations. Since *WordNet* or *WOLF* do not relate words but senses, each node is split into as many senses as the

concept has, and each edge connects exactly two senses. In essence, if a word has *n* possible senses, it will initially have *n* different lexical chain links associated with it. Afterwards, when adding a new lexical chain link to the disambiguation graph, new connections need to be added between the concept and all the other related links in the graph.

The types of semantic relations taken into consideration when linking two words are hypernymy, hyponymy, synonymy, antonymy, or whether the words are siblings by sharing a common hypernym. The weights associated with each relation vary according to the strength of the relation and the proximity of the two words in the text analyzed. Table 6 depicts the weights later used in *ReaderBench* (see *7 ReaderBench* (1) – Cohesion-based Discourse Analysis and Dialogism), similar to Galley and McKeown (2003), but with antonymy having importance (and associated weights) equivalent to the synonymy relation.

Table 6. Lexical chains – adapted weights based on semantic relations and word distances (after Galley & McKeown, 2003).

	Distance between words			
Semantic relations	1 sentence	3 sentences	same block/paragraph	other
Synonym/Antonym	1	1	.5	.5
Hypernym/Hyponym	1	.5	.3	.3
Sibling	1	.3	.2	0

The pruning of the disambiguation graph corresponds to the actual disambiguation step of the algorithm (Galley & McKeown, 2003). Therefore, for each word, the values of the lexical chain links associated with each of the word senses are compared and the link with the best value is selected. The value of a link is computed as the sum of the weights of all the connections for that link or, in terms of the generated graph, the sum of the weights of all the edges connecting that link to other links in the graph. In the end, when a specific word is associated to a link, it has been disambiguated. The last step consists of removing all other links associated with the word's other senses, from the disambiguation graph. In order to optimize the process of identifying the link with the best value, these values can be computed incrementally when building the disambiguation graph, as new connection between two links are added. In this particular context, each lexical chain is, in fact, in itself a graph or, to be more exact, a connected component of the pruned disambiguation graph. Therefore, lexical chains are identified as connected components within the disambiguation graph by using the breadth-first search algorithm (Cormen et al., 2009).

4.3.2 Semantic Similarity through Tagged LSA

Latent Semantic Analysis (LSA) (Deerwester et al., 1989; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997; Dumais, 2004) is a natural language processing technique starting from a vector-space representation of semantics highlighting the co-occurrence relations between terms and containing documents, after that projecting the terms in sets of concepts (semantic spaces) related to the initial texts. LSA builds the vector-space model, later on used also for evaluating similarity between terms and documents, now indirectly linked through concepts (Landauer, Foltz, et al., 1998; Manning & Schütze, 1999). Moreover, LSA can be considered a mathematical method for representing words' and passages' meaning by analyzing in an unsupervised manner a representative corpus of natural language texts. More formally, LSA uses a sparse term-document matrix that describes the occurrence of terms in corresponding documents. LSA performs a "bag-of-words" approach as it disregards word order by counting only term occurrences, later to be normalized. The indirect link induced between groups of terms and documents is obtained through a singular-value decomposition (SVD) (Golub & Kahan, 1965; Golub & Reinsch, 1970; Landauer, Laham, & Foltz, 1998) of the matrix, followed by a reduction of its dimensionality by applying a projection over *k* predefined dimensions, similar to the least-squares method (see Figure 17).

From a cognitive point of view, LSA has been thoroughly analyzed, with two prominent directions. Firstly, LSA can be seen as an expression of meaning as each word can be represented as a context-free vector in the semantic vector-space model (Kintsch, 2000, 2001). The actual dimensions of concepts do not bear a specific individual meaning, but the overall representation generated by LSA can be considered a map of meanings (Landauer et al., 2007). Secondly, the semantic proximity effect (Howard & Kahana, 1999) highlights the positive correlation between the similarities measured through LSA and the human recall using word association lists. Moreover, it was noted that the inter-response time between similar words was much quicker than for dissimilar words, justifying that LSA bears resemblance to the human memory, more specifically to memory search and free recall (Zaromb et al., 2006; Landauer et al., 2007). Also, the evolution modeled through increasing corpora dimensions for deducing the word maturity metric (Landauer et al., 2011) underpins the cognitive similarities of word associations in terms of prior information or knowledge.

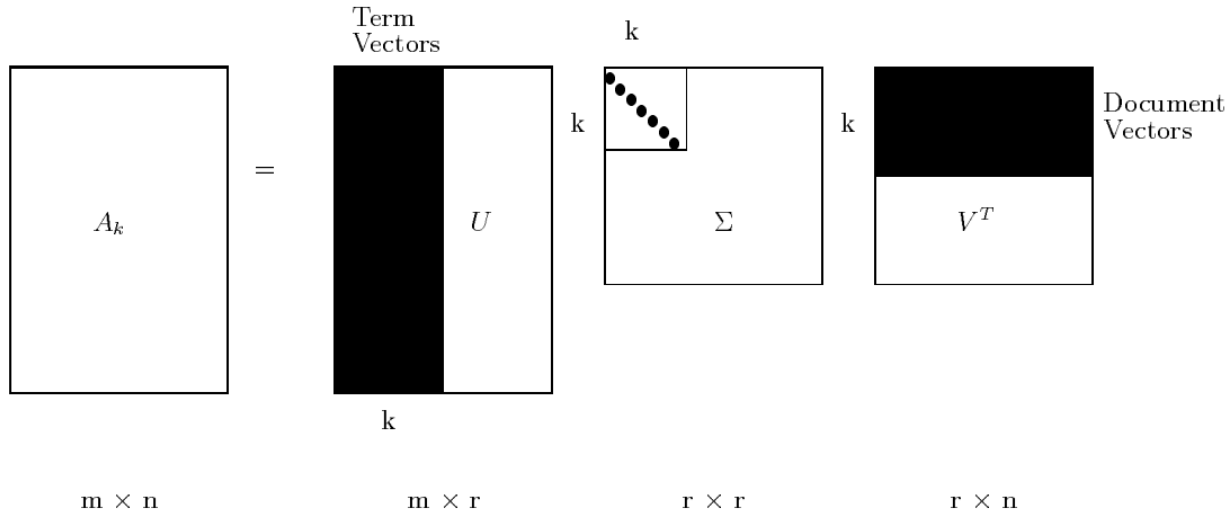


Figure 17. Latent Semantic Analysis Decomposition (Berry, Dumais, & O'Brien, 1995, p. 5).

From a computational perspective, LSA is used for evaluating the proximity between concepts or textual elements by cosine similarity or, equivalent, scalar product (see Equation 1). In addition to the initial model, multiple optimizations can be envisioned in order to increase the reliability of the semantic vector-space. Firstly, two crucial aspects, although empirical, need to be addressed: the initial document dimension and the number of dimension *k* after projection. In terms of documents size, semantically and topically coherent passages of approximately 50 to 100 words are the optimal units to be taken into consideration while building the initial matrix (Landauer & Dumais, 2008). While considering the number of dimensions *k*, 300 can be considered an optimal empiric value agreed by multiple sources (Berry, Drmac, & Jessup, 1999; Jessup & Martin, 2001; Lizza & Sartoretto, 2001; Landauer et al., 2007; Lemaire, 2009).

$$Sim(word_1, word_2) = \frac{\sum_{i=1}^{k} word_{1,i} * word_{2,i}}{\sqrt{\sum_{i=1}^{k} word_{1,i}^2} * \sqrt{\sum_{i=1}^{k} word_{2,i}^2}} \quad (1)$$

Secondly, term weighting (Dumais, 1991) can be applied on the elements of the initial term-document matrix. Term frequency – inverse document frequency (*Tf-Idf*) (Manning & Schütze, 1999) provides a practical approach due to its duality: 1/ local importance, reflected in the normalization of the number of appearances of a word in a given document and 2/ global significance by weighting the appearances of a given word in all corpus documents, therefore enhancing the importance of rare words and reducing the significance of common ones (see Equation 2). Moreover, although word vectors can be directly summed up in order to build the representation of larger textual fragments, normalization of contained concepts also improves overall performance.

$$w_{D,i} = \left(\ln\left(tf_{D,i} + 1\right)\right) * ln\frac{N}{n_i} \quad (2)$$

where $tf_{D,i}$ is the number of occurrences of the term i in document D, N is the total number of documents in the corpus and n_i is the number of documents in which the term i occurs.

Thirdly, POS tagging (Wiemer-Hastings & Zipitria, 2000; Rishel, Perkins, Yenduri, & Zand, 2006) can be applied on all remaining words after stop word elimination and all inflected forms can be reduced to their lemma (Dascalu, Trausan-Matu, et al., 2010b; Bestgen, 2012), that means enforcing the NLP pipe on the training corpus. According to Lemaire (2009) and Wiemer-Hastings and Zipitria (2000), stemming applied on all words reduces overall performance because each inflected form can expresses different perceptions and is related to different concepts. Therefore, as compromise of all previous NLP specific treatments, the latest version of the implemented tagged LSA model (Dascalu, Dessus, et al., in press; Dascalu, Trausan-Matu, et al., in press) uses lemmas plus their corresponding part-of-speech, after initial input cleaning and stop words elimination. In the end, due to the high demand of computational resources when performing the SVD decomposition on a sparse matrix of at least 20K terms with 20K passages (Landauer & Dumais, 2008) (see *7.2* Cohesion-based Discourse Analysis), distributed computing enabling a concurrent and parallel execution of tasks can be considered a necessity for increasing speedup.

Similar to semantic distances, we must also consider the limitations of LSA, correlated to the experiments performed by Gamallo and Bordag (2011): 1/ the requirement of a large corpus of documents for training, both domain specific and general; 2/ the computational constraints due to the SVD decomposition phase; 3/ the model is blind to word order and to polysemy, as all word senses are merged into a single concept; 4/ the empirical selection of *k* and the segmentation of the initial documents into cohesive units of a given size, although co-occurrence patterns emerge in large training corpora; and 5/ despite the fact that updating mechanisms have been devised for increasing the training corpora (Berry et al., 1995; Witter & Berry, 1998), it is unfeasible to apply them in practice, and once trained, the model remains unchanged.

4.3.3 Topic Relatedness through Latent Dirichlet Allocation

The goal of Latent Dirichlet Allocation (LDA) topic models is to provide an inference mechanism of underlying topic structures through a generative probabilistic process (Blei et al., 2003). Starting

from the presumption that documents integrate multiple topics, each document can now be considered a random mixture of corpus-wide topics (see Figure 18). In order to avoid confusion, an important aspect needs to be addressed: topics within LDA are latent classes, in which every word has a given probability, whereas topics that are identified within subsequently developed systems (*A.S.A.P.*, *Ch.A.M.P.*, *PolyCAFe* and *ReaderBench*) are key concepts from the text. Additionally, similar to LSA, LDA also uses the implicit assumption of the bag of words approach that the order of words doesn't matter when extracting key concepts and similarities of concepts through co-occurrences within a large corpus. In contrast to LSA (Landauer, 2002) and *WordNet* (Miller, 1998) that have empirically proved cognitive bases, LDA does not have such a cognitive argumentation; it is a probabilistic topic model in which the connotations of the latent space behind the model can be ignored (J. Chang, Boyd-Graber, Wang, Gerrish, & Blei, 2009).

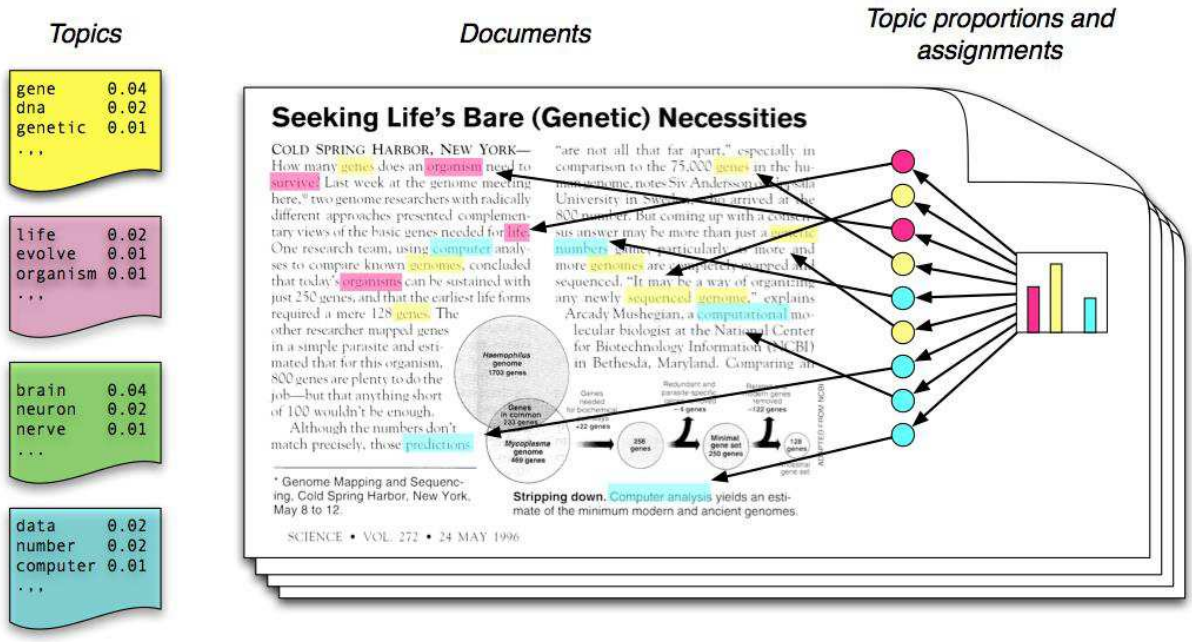


Figure 18. Latent Dirichlet Allocation – visualization of underlying variables (Blei, 2012, p. 78). Including observed words (the actual terms from the initial document), topic classes, per word topic assignments (circles for each concept, with specific class encoding corresponding to the term's dominant topic) and per document proportions (the distribution of the 4 topics depicted as a bar diagram)

Every topic contains a probability for every word, but after the inference phase a remarkable demarcation can be observed between salient or dominant concepts of a topic and all other vocabulary words. In other words, the goal of LDA is to reflect the thematic structure of a document or of a collection through hidden variables and to infer this hidden structure by using a posterior inference model (Blei et al., 2003) (see Figure 19). Later on, as documents can be considered a

mixture of topics, LDA focuses on situating new documents in the estimated pre-trained model. A topic is a Dirichlet distribution (Kotz, Balakrishnan, & Johnson, 2000) over the vocabulary simplex (the space of all possible distributions of words from the training corpora) in which thematically related terms have similar probabilities of occurrences. Moreover, as the Dirichlet parameter can be used to control sparsity, penalizing a document for using multiple topics, LDA's topics reflect in the end sets of concepts that co-occur more frequently (Blei & Lafferty, 2009).

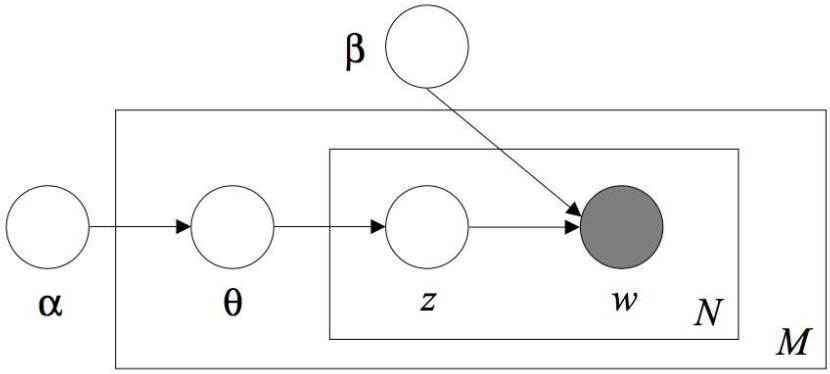


Figure 19. Latent Dirichlet Allocation – graphical model representation (Blei et al., 2003).
$w_{d,n}$ – n^{th} observed word in d document; $z_{d,n}$ – per word topic assignment; θ_d – per document topic proportions; β_k – per corpus topics distributions; M – corpus of documents; α – Dirichlet parameter; Each structure can be considered a random variable

Therefore, documents become topics distributions drawn from Dirichlet distributions and similarities between textual fragments can be expressed by comparing the posterior topic distributions. Due to the fact that KL divergence (see Equation 3) (Kullback & Leibler, 1951) is not a proper distance measure, as it is not symmetric, Jensen-Shannon dissimilarity (see Equation 4) (Manning & Schütze, 1999; Cha, 2007) can be used as a smoothed, symmetrized alternative. In the end, semantic similarity between textual fragments can be computed in terms of relatedness between distributions of topics – $prob(fragment_i)$, more specifically the inverse of the Jensen-Shannon distance (see Equation 5):

$$D_{KL}(P||Q) = \sum_i \left(\frac{P(i)}{Q(i)}\right) P(i) \quad (3)$$

$$D_{JS}(P||Q) = \frac{1}{2}(D_{KL}(P||M) + D_{KL}(Q||M)), M = \frac{1}{2}(P + M) \quad (4)$$

$$sim(fragment_1, fragment_2) = 1 - D_{JS}(prob(fragment_1), prob(fragment_2)) \quad (5)$$

Despite the fact that LDA uses only few latent variables, exact inference is generally intractable (Heinrich, 2008). Therefore, the solution consists of using approximate inference algorithms, from which Gibbs sampling (Griffiths, 2002) seems most appropriate and is most frequently used. Gibbs sampling can be considered a special case of Markov-chain Monte Carlo (MCMC) simulation (MacKay, 2003) and integrates relatively simple algorithms for approximating inference in high-dimensional models (Heinrich, 2008) – *k*, the number of topics, is usually 100, as suggested by Blei et al. (2003). Of particular interest from a computational point of view is the possibility to perform a distributed Gibbs sampling (McCallum, 2002) in order to increase training speedup.

Although LDA proved to be reliable in extracting topics and has the lowest perplexity levels (a measure algebraically equivalent to the inverse of the geometric mean per-word likelihood) when compared to other probabilistic semantic models (Blei et al., 2003), we must also consider its drawbacks: 1/ although topics reflect terms that more tightly co-occur, there are no actual class significances automatically deduced and topics are not equi-probable (Arora & Ravindran, 2008); 2/ by using an approximate inference model, there are inevitably estimation errors, more notable when addressing smaller documents or texts with a wider spread of concepts, as the mixture of topics becomes more uncertain; 3/ similarly to LSA, LDA is blind to word order, but polysemy is reflected in the membership of the same word, with high probabilities, in multiple topics; and 4/ LDA, in comparison to LSA, loses the cognitive significance and the posterior distributions are nevertheless harder to interpret than the semantic vector space representations of concepts.

Part 2 – Empirical Studies

Overview of Empirical Studies

Across time, a series of systems were developed in our research group for analyzing discourse in CSCL sessions (see Table 7 for some of them). Each of them integrated additional facilities and a more detailed and comprehensive perspective starting from the previously conducted evaluations (see Table 7). *A.S.A.P.* and *Ch.A.M.P.* were the first developed systems before the commencement of this thesis, which addressed in extent participant grading and provided valuable insight, traceability and continuity in terms of the evolution of the inter-connected functionalities. Special attention must be addressed to *Polyphony* that significantly broadened the perspective in terms of chat analysis, with emphasis on feedback generation for both learners and tutors, and whose main ideas, general architecture and evaluation framework were elaborated in *ReaderBench*.

Table 7. Main developed systems and the evolution in time of their purposes.

System name	Period	Main purposes	Principal references
Polyphony	2007	Initial evaluations of chat conversations introducing the polyphonic perspective	(Trausan-Matu, Rebedea, et al., 2007; Trausan-Matu & Stahl, 2007)
Advanced System for Assessing Chat Participants (*ASAP*)	2008	Preliminary chat analysis in terms of quantitative participant involvement (mostly SNA)	(Dascalu, Chioasca, & Trausan-Matu, 2008b; Dascalu et al., 2008a)
Chat Assessment and Modeling Program (*Ch.A.M.P.*)	2009	Diversity of evaluation factors Automatic weighting of factors	(Dascalu & Trausan-Matu, 2009b, 2009a; Dascalu, Trausan-Matu, et al., 2010b)
Polyphonic Conversation Analysis and	2009 – 2011	Comprehensive chat analysis and feedback generation	(Rebedea et al., 2010; Trausan-Matu, Dessus, et al., 2010; Trausan-Matu & Rebedea, 2010; Dascalu, Rebedea, et al., 2011; Rebedea, Dascalu, Trausan-Matu, Armitt, et al., 2011; Trausan-Matu et

Overview of Empirical Studies

System name	Period	Main purposes	Principal references
Feedback (*PolyCAFe*)		Refined natural language processing techniques and discourse analysis	al., 2011; Trausan-Matu, Dascalu, & Rebedea, 2012)
		Collaboration assessment	
		Distributed computing facilities	(Dascalu, Dobre, Trausan-Matu, & Cristea, 2011)
ReaderBench	2012 – present	Discourse structure – Cohesion Graph	(Trausan-Matu, Dascalu, & Dessus, 2012; Dascalu, Dessus, et al., in press)
		Multi-dimensional assessment of textual complexity	(Dascalu et al., 2012)
		Automatic identification of reading strategies	(Dessus et al., 2012; Dascalu, Dessus, et al., in press; Oprescu, Dascalu, Trausan-Matu, Dessus, & Bianco, in press)
		Collaboration assessment	(Dascalu, Trausan-Matu, et al., in press)

Besides the actual evolution and the integration of additional facilities and tools (see Table 8), of particular interest is the cognitive and educational trend that has been progressively followed. In addition to this general perspective, we have introduced a traceability matrix of covered functionalities and of integrated tools before describing in detail each system, in its corresponding dedicated section. Although the entire initial research (*Polyphony*, *A.S.A.P.*, *Ch.A.M.P.* and *PolyCAFe*) was mainly focused on chat analysis, a major discrepancy can be identified among the systems. Whereas the *A.S.A.P.* and *Ch.A.M.P.* were focused on a quantitative evaluation of participant's involvement and their main purpose was to provide a final grade to each chat participant, *PolyCAFe*, similar to some extent to the initial *Polyphony* system, changed the perspective, offering a dual view – learner and tutor oriented – designed to enhance the collaborative learning experience through the provided feedback.

Table 8. Comparison of provided features and tools across the developed systems.

Category	Functionality	*Polyphony*	*ASAP*	*ChAMP*	*PolyCAFe*	*ReaderBench*
Underlying discourse structure	Utterance graph	✓*	✓*	✓*	✓	
	Cohesion graph					✓
	Speech acts identification	✓*	✓*	✓*	✓	
	Topics modeling	✓*	✓*	✓*	✓	✓
	Dialogical perspective	✓			✓	✓

Overview of Empirical Studies

Category	Functionality	*Polyphony*	*ASAP*	*ChAMP*	*PolyCAFe*	*ReaderBench*
Individual learning	Reading strategies identification					✓
	Textual complexity assessment					✓
Collaborative learning	Participant involvement evaluation	✓*	✓*	✓	✓	✓
	Intervention scoring	✓*	✓*	✓	✓	✓
	Conversation visualization	✓			✓	✓
	Collaboration assessment				✓	✓
	Comprehensive feedback delivery				✓	
	Semantic search				✓	
	Semantic extractive summarization				✓	✓
Generic	Multilingual support					✓
Tools	Patterns	✓*	✓*	✓*	✓	✓
	Information Retrieval techniques (e.g., *Tf-Idf*)		✓*	✓*	✓	✓
	NLP pipe		✓*	✓	✓	✓
	Semantic distances in ontologies	✓*			✓	✓
	Latent Semantic Analysis (LSA)		✓*	✓	✓	✓
	Latent Dirichlet Allocation (LDA)					✓
	Genetic Algorithms (GA)			✓		
	Support Vector Machines (SVM)					✓
	Social Network Analysis (SNA)		✓	✓	✓	✓
	Distributed computing				✓	✓

* – partial support

Additionally, although not specifically stated, the initial tendency was to replace the tutor and to solely provide in the end automatic grades for each participant. In this direction, multiple factors were combined using an automatic weight optimization algorithm, trying to best match the assigned student grades (the scores generated by the system, on one hand, and the tutor's mark, on the other). After performing the initial evaluations using *A.S.A.P.* and *Ch.A.M.P.,* this approach turned out to be not as feasible as expected, mostly in terms of its implications in the educational scenario in which the conversations took place. For learners, the final grade has much more significance with a related description or explanation. Without a comprehensive presentation of the presented factors, its significance dramatically decreased.

From the tutor perspective, although promising, the lack of aggregation and the mere exhaustive presentation of all evaluation factors seemed tiresome and, in the end, full trust could not be committed in directly using the outputs of the systems, without personally inspecting the contents of each conversation. Therefore, without any cognitive benefits and a doubling of effort (both automatic and manual, but parallel and unsupportive one over the other), a shift had to be made towards providing support and comprehensive feedback to both learners and tutors. On the other hand, *PolyCAFe*, that can be a considered a successor and an initial integrator of *Polyphony*, *A.S.A.P.* and *Ch.A.M.P.* at macroscopic level, aimed to extract significant information from chat analysis in order to provide feedback to both learners and tutors.

This change of perspectives can be also clearly seen in the evolution of used indicators. At the beginning, the goal of *A.S.A.P.* was to best fit the automatic scoring process for minimizing the overall error. Later on, in order to decrease bias in the evaluation, an increased correlation with the tutor grades was sought, obtainable through the weight optimization algorithm employed in *Ch.A.M.P.*. Afterwards, when considering the validation process of *PolyCAFE*, the educational dimension played a central role: relevance, quality and consistency of the provided feedback or measurements of the reduction in time required for manual assessment, were just some of the evaluated metrics.

Nevertheless, *A.S.A.P.* and *Ch.A.M.P.* provided a strong basis for later development, a wide variety of evaluation factors, but also highlighted the minuses of the involvement-centered, quantitative approach. Convergence towards *PolyCAFe* was achieved by combining the prior technical knowledge

with an educational perspective that dramatically increased the usefulness of the developed educational platform.

In contrast to the previous systems, *ReaderBench* can be considered a different "*species*" as the analysis now covers, besides a refined chat and collaboration analysis, general texts as an extension to reading materials, meta-cognitions evaluation through the identification of reading strategies from learner verbalizations and textual complexity assessment. By considering the underlying cohesion graph and the proposed voice analysis, *ReaderBench* provides in-depth support to learners and tutors through the wide variety of functionalities it integrates. Now, emphasis is put on comprehension, cognitive modeling induced by higher-level reading strategies (e.g., bridging), on the synergic effect of voice overlapping and inter-animation, but also on the social knowledge-building effect generated through collaboration. In this context, *ReaderBench* covers a larger mixture of learning activities, envisions complex educational scenarios and continues the trend in providing useful feedback to both tutors and learners.

The next chapter is focused on presenting in detail *A.S.A.P.* and *Ch.A.M.P.* that address in extent the assessment of participants' involvement in multi-party chat conversations and the proposed distributed computing architecture applicable for any evaluation model, that was later on enforced on both *PolyCAFe* and *ReaderBench*. Altogether, these approaches represented the starting point of this thesis and provided valuable insight for the work carried on subsequently. Afterwards, we make a detailed presentation of the core systems, *PolyCAFe* and *ReaderBench*, with emphasis on main functionalities, improvements in comparison to previous versions, relevant interfaces and validation results.

5 Quantitative Analysis of Chat Participants' Involvement

5.1 *A.S.A.P.* – Advanced System for Assessing Chat Participants

5.1.1 General Presentation

The first experiments were performed with *A.S.A.P.* (Dascalu et al., 2008a, 2008b) (see Table 9) whose purpose was to discover the most competent user in a chat using several analysis factors, starting with the simplest, such as the dimension of utterances, and ending with pragmatics issues such as speech acts (Austin, 1962; Searle, 1969; Trausan-Matu, Chiru, & Bogdan, 2004) and even social aspects of interaction between conversation participants (see *3.2* Social Network Analysis). The initial educational scenarios consisted on carrying out multi-participant chat conversations in an academic environment in which students had to debate on given topics, pre-specified by the tutor. Afterwards, their conversations were evaluated in the first iteration through a peer review process, by their colleagues, and afterwards by their tutor, as the initial assessments were prone to bias and high discrepancies between evaluations were encountered.

Table 9. *A.S.A.P.* Traceability matrix of provided functionalities and integrated tools.

	Tools				
Functionality	Patterns	IR	NLP pipe	LSA	SNA
Utterance graph	✓*			✓*	
Speech acts identification	✓*				
Topics modeling		✓*		✓*	
Participant involvement evaluation		✓*	✓*	✓*	✓
Intervention scoring		✓*	✓*		

* – partial support

Taking as a starting point the utterances of a chat, their sequencing and the social network of the participants, *A.S.A.P.* (see Figure 20) was designed to assign a grade to each participant reflecting his involvement throughout the discussion. The system was conceived to provide a complementary approach to the one of *Polyphony* (Trausan-Matu, Rebedea, et al., 2007), the first system developed within our research group, in the sense that its aim was to integrate in-detail metrics for determining each participant's involvement throughout a chat conversation. The main user interface (see Figure 20) presents: the user rankings, the main concepts/topics automatically extracted, events generated while processing the conversation and multiple tabs for each evaluation factor. With regards, to topics extraction, the approach was rather straightforward, based only on term frequency – inverse document frequency (*Tf-Idf*) (Manning & Schütze, 1999).

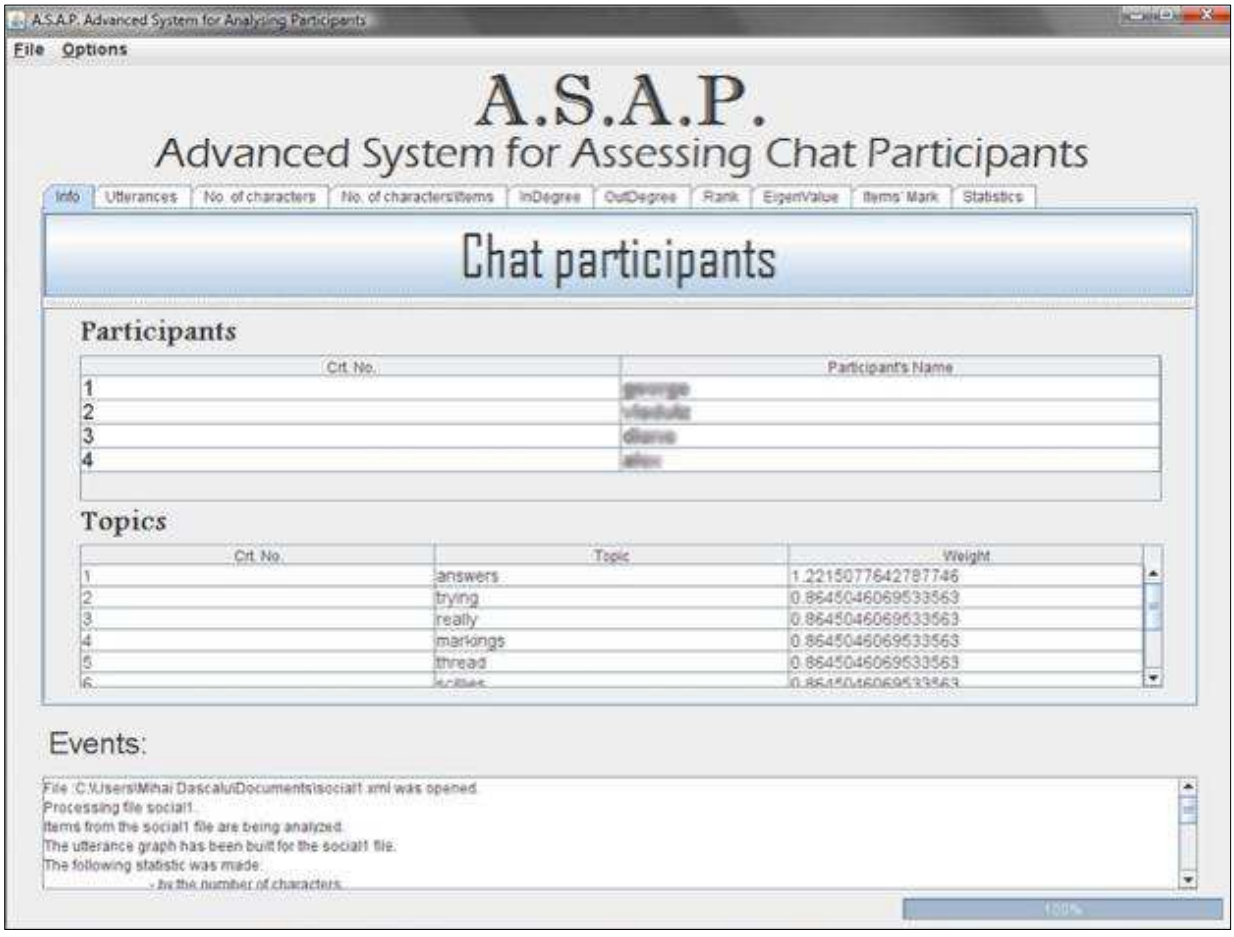


Figure 20. *A.S.A.P.* Main user interface.

When considering the used factors, the number of written characters is probably the simplest feature to take into account when searching for the most competent participant in a chat conversation. This feature provides an indicator of each participant's contribution, but it is obviously naive because the intervention may be off-topic. In this context, it can be used in conjunction with the number of

utterances as an indicator of a talkative user, which, in most cases, is different from the knowledgeable one. Therefore, in addition to simple statistics at character level (e.g., average number of characters per utterance), features that reflect the comprised content should influence the final participant scores. For this purpose, even the first analysis process considered the keywords or the topics of the discussion, in this case predefined by the tutor.

Moreover, besides quantity and, in some extent, quality seen as predefined topics coverage inferred from the utterances, social factors were also taken into account. Consequently, an utterance graph (Trausan-Matu, Rebedea, et al., 2007) is generated from the chat transcript in concordance with the utterances exchanged between the participants. Starting from the graph theory (Cormen et al., 2009) the number and the distribution of edges the participants' social network is analyzed providing a good estimator of users' involvement within the conversation. In our specific case, the *interaction graph* considers as nodes the participants of the conversation and as edges the set of inter-changed interventions between each pair of speakers, linked through explicit or implicit links. The *explicit links* between utterances are generated through the referencing facility of chat systems as *ConcertChat* (Holmer et al., 2006), whereas the *implicit links* are discovered through natural language processing techniques (Trausan-Matu et al., 2004; Stahl, 2006b; Rebedea, 2012) (see *4.3* Natural Language Processing Techniques). As an extension of the classic interaction model that considers the number of inter-changed utterances between different participants, we can also envisage that edges of this social network are reflected by a cumulative function applied on a scoring mechanism for each intervention.

From the graph theory's point of view, the first two measures taken into consideration for the subsequent processing of the participants' social network (interaction graph) and of the utterance graph (see Figure 21) are in-degree and out-degree (Brandes, 2001). In addition, multiple centralities (e.g., closeness, graph centrality, betweenness, stress, or eigenvector) (Freeman, 1977; Brandes, 2001) and an adaptation of the Google Page Rank algorithm (L. Page, 2001) deriving the user rank were also used to better express the participants' involvement and his/her importance within the conversation.

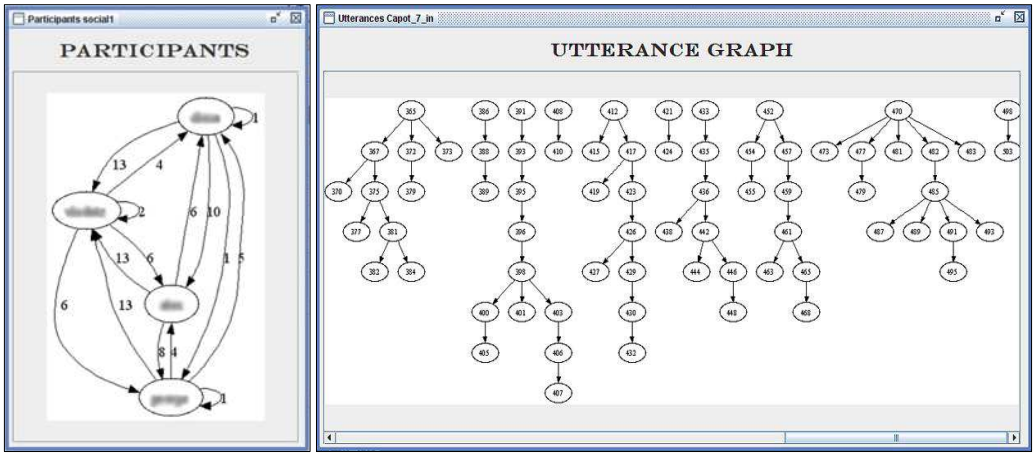


Figure 21. *A.S.A.P.* Participants' social network and the utterance graph.

Starting from the previous analysis of social interactions, a first taxonomy of users was also sketched comprising the following categorization of chat participants: *knowledgeable*, *gregarious*, *passive* and *inactive*. From the social network's point of view, a *knowledgeable* user is one who most influences others with whom he/she is directly linked. He is usually defined as a participant who receives a lot of questions and answers them correctly in a large proportion and, therefore, is more sought-after. On the other hand, a *gregarious* user is a very active participant in the chat, enlarging the question pool. If one is to analyze the activity of groups, gregarious users tend to have a high out-degree, while the experts – usually marked as knowledgeable users – emerge due to a high in-degree score, their activity being crucial for the overall consistency of the conversation.

A *passive* user generally responds only to questions specifically addressed to him. Because of the lack of involvement in active threads, although his remarks might have an important role to play in the overall evolution of the chat, the evaluation system had difficulties estimating his/her overall performance, mostly from the perspective of social network analysis. This type must not be mistaken for an *inactive* user who has no direct impact on the ongoing discussion, posing no important questions, replies or remarks in the chat. The differentiation between the latter two was performed by comparing the average utterance scores, per participant.

Although most participants are a mixture of multiple categories or change their behavior during a longer conversation, an initial classification was useful for pinpointing different traits (see Table 10).

Table 10. *A.S.A.P.* Participant taxonomy.

		Involvement	
		Low	High
Proven domain knowledge/topics coverage	Low	Inactive	Gregarious
	High	Passive	Knowledgeable

In order to perform the previous classification, the first scoring mechanism of participants' interventions was proposed, that considered, besides the length of an utterance, the following factors:

- The number of keywords which remain after the spell-check, stemming and stop words elimination (Manning & Schütze, 1999).
- Normalized word occurrences (Jurafsky & Martin, 2009).
- The level at which the participant's intervention can be found in the utterance graph (see Figure 21) that was constructed using the explicit links provided by the referencing facility (Holmer et al., 2006).

In addition to evaluating each intervention and assigning a corresponding score, two dimensions were of particular interest: 1/ the visualization of a discussion thread identified at this point by applying a breadth first algorithm (Cormen et al., 2009) from a given utterance and by using the transposed explicit links as edges (see Figure 22.a) and 2/ an overall view of the conversation from which zones with higher scores, therefore with a higher importance or relevance, could be identified (see Figure 22.b). In addition, the score of a discussion thread may be raised or lowered by each utterance (see Figure 22.a) as the displayed score takes into account the previously cumulated value, the score of the current utterance and a positive or negative weight, empirically established in terms of the identified speech acts (Austin, 1962; Searle, 1969; Trausan-Matu et al., 2004). Heuristics and cue phrases were used (Trausan-Matu et al., 2004) for the identification of speech acts, whereas the weights were experimentally set to reflect different succession scenarios (e.g., question followed by a negation, declaration followed by confirmation) with values in the [-1.5; 1.5] interval. This approach was later on discarded as it was not extensible in terms of newly identified speech acts and due to multiple conflicting situations that arose during evaluations, as multiple speech acts were identified within a single utterance and it became rather complicated to choose from the multitude of predefined values.

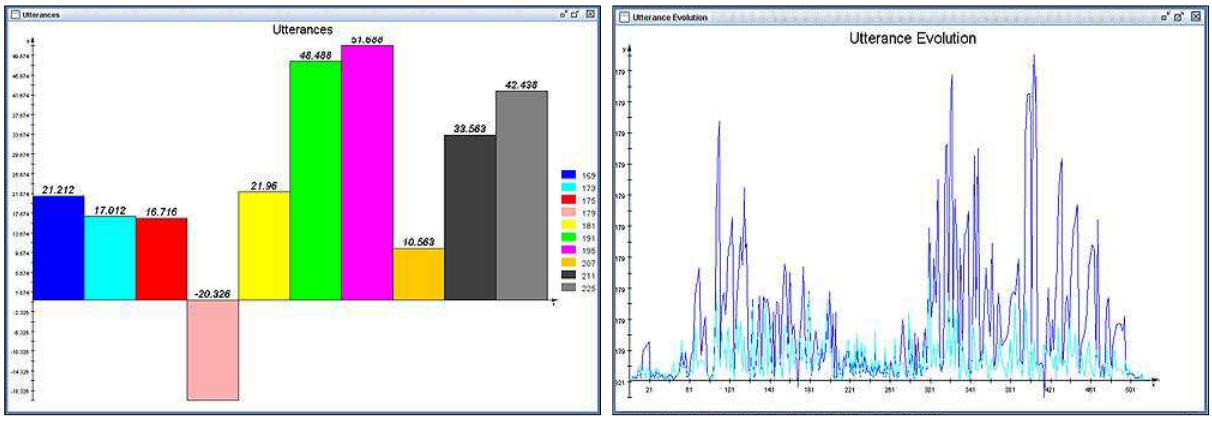


Figure 22. *A.S.A.P.* Charts depicting: a. Utterances' evolution in a single explicit thread; b. Utterances' score evolution during the entire conversation.

The utterance scores are combined per speaker into an overall participant score and are used for building a classification and an evolution in time of participants throughout the ongoing conversation (see Figure 23.a). In addition, starting from each SNA factor applied on the interaction graph, individual charts are generated, presenting a comparative view of the participants' involvement in the overall discussion based on that specific factor (see Figure 23.b). In the end, each participant received a grade on a [0; 10] scale, linearly distributed in terms of the minimum and maximum scores of all participants involved in the discussion.

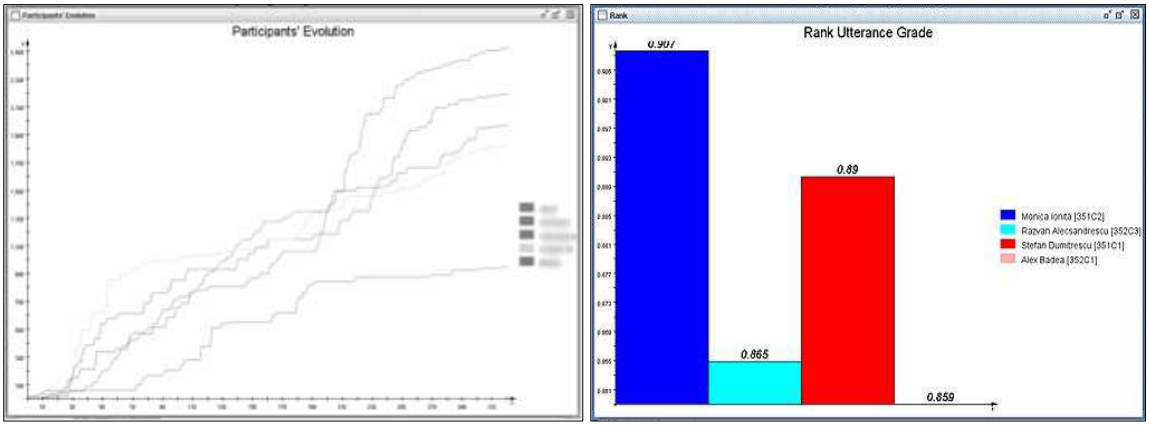


Figure 23. *A.S.A.P.* Charts representing: a. Overall participants' evolution; b. Comparative results of participants for a given SNA factor applied on the interaction graph.

5.1.2 Annotation Tool

Annotated corpora are needed in order to run machine-learning algorithms that train the analysis modules and to evaluate and fine-tune the assessment tool. Therefore, in order to facilitate the annotation, an editing program (*C.An.* – Chat Annotator) was developed (see Figure 24), later used

to build the gold standard for evaluating the performance of each developed system. In the end, the tool was used by more than 250 students in two courses organized at the University Politehnica of Bucharest between 2008 and 2010: Human-Computer Interaction (4th year bachelor students) and Natural Language Processing (5th year license students or 1st year master students).

Figure 24. *C.An.* Chat Annotator.

The application allows the annotation of chat conversations in an XML format with the following information:

- Comments or grades for each individual utterance.
- Grades for a participant's 20 successive utterances, reflecting his/her contribution to that segment of the conversation.
- Global grades for each participant in a given conversation, with emphasis on the involvement and collaboration dimensions.
- Two global grades for the discussion as a whole: the degree at which collaboration was successfully achieved and the degree to which the conversation follows the tutor-defined topics of the discussion (on topic relatedness).
- Annotation of implicit links between utterances, including their type and associated patterns.
- Identification of the main topics (keywords) of the conversation.
- Segments of successful collaboration denoted as intense collaboration zones.

5.1.3 Preliminary Validation of *A.S.A.P.*

An initial experiment was performed on a corpus of 32 conversations that were each manually assessed by at least a student colleague. As background, all involved participants were students in the 4th year or in the 1st year of master in the domain of computer science, and they had to debate the use, the advantages and the disadvantages, of different CSCL technologies. As we were dealing with a time-consuming process, we opted to conduct these initial experiments using a peer-based assessment in order to cope with the feasibility and validity of the automatic chat analysis process. Although focusing mostly on a quantitative evaluation, the results were promising – an average error of 2.5 on a [0; 10] scale of grading, while considering the absolute value of the differences between the automatic and the annotated scores.

Unfortunately, this initial evaluation was biased as we identified major discrepancies in terms of the grades assigned by students, even for the same conversation, but especially between different ones, as there was no unitary baseline of evaluation. Moreover, Pearson correlations could not be used on this initial corpus as multiple conversations had received a universal grade for all participants. Therefore, all subsequent experiments used the grades assigned by or verified by tutors in order to limit these drawbacks.

Nevertheless, as *A.S.A.P.* was rather close to the students' assessment of involvement, this initial research gave us a strong incentive that with further improvements, the automatic grading process can become a viable alternative to the manual assessment. Nevertheless, the omnipresent restriction that we are dealing with the human, subjective factor within the manual evaluation must be also taken into account.

5.2 *Ch.A.M.P.* – Chat Assessment and Modeling Program

5.2.1 General Presentation

With the continuous evolution of collaborative environments, the needs of automatic analyses of participants in instant messenger discussions or conferences have become essential. For this aim, a new system was developed that proposes an integrated scoring mechanism for the evaluation of participants' involvement. In this context, a series of factors for thoroughly assessing participants were taken into consideration: measures derived from Page's essay grading techniques (E. Page, 1966;

Wresch, 1993), readability formulas (Brown, 1998), social network analysis metrics (Mislove, Marcon, Gummadi, Druschel, & Bhattacharjee, 2007; Newman, 2010) (see *3.2* Social Network Analysis), natural language processing (including lexical analysis) (Manning & Schütze, 1999), Latent Semantic Analysis (Landauer, Foltz, et al., 1998) (see *4.3* Natural Language Processing Techniques) and data-mining techniques (Jurafsky & Martin, 2009) (see Table 11).

Table 11. *Ch.A.M.P.* Traceability matrix of provided functionalities and integrated tools.

	Tools					
Functionality	Patterns	IR	NLP pipe	LSA	GA	SNA
Utterance graph	✓*			✓		
Speech acts identification	✓*					
Topics modeling		✓*		✓		
Participant involvement evaluation		✓*	✓	✓	✓	✓
Intervention scoring		✓	✓	✓		

* – partial support

The weights of each factor in the overall scoring mechanism are optimized using a genetic algorithm whose entries are provided by a perceptron in order to ensure numerical stability. Since the mechanics were completely rebuilt, the dimensions of the analysis in terms of identified factors were fundamentally changed and multiple visualizations were added to the interface, we opted for changing the system's name (Dascalu & Trausan-Matu, 2009b, 2009a) (see Figure 25). In comparison to *A.S.A.P.* (see Figure 20), accent was put on usability: the main user interface has been simplified, the interaction graph is displayed as physical and radial models using Prefuse (Heer et al., 2005) (http://prefuse.org/) and the evaluation factors are grouped into relevant categories.

5.2.2 The Scoring Process

Communication between participants in a chat is conveyed through language in a written form. Lexical, syntactic, and semantic information are the three levels used to describe the features of written utterances (Anderson, 1985), and all were taken into account within *Ch.A.M.P.* in order to analyze a participant's involvement in a chat. First, surface metrics are computed for all the utterances of a participant in order to determine factors like fluency, spelling, diction or utterance structure (E.

Page, 1966). All these factors are linearly combined and a score is obtained for each participant without taking into consideration a lexical or a semantic analysis of what they are actually discussing. At the same level, readability ease measures expressed by simple readability formulas are computed (Davison & Kantor, 1982; Brown, 1998).

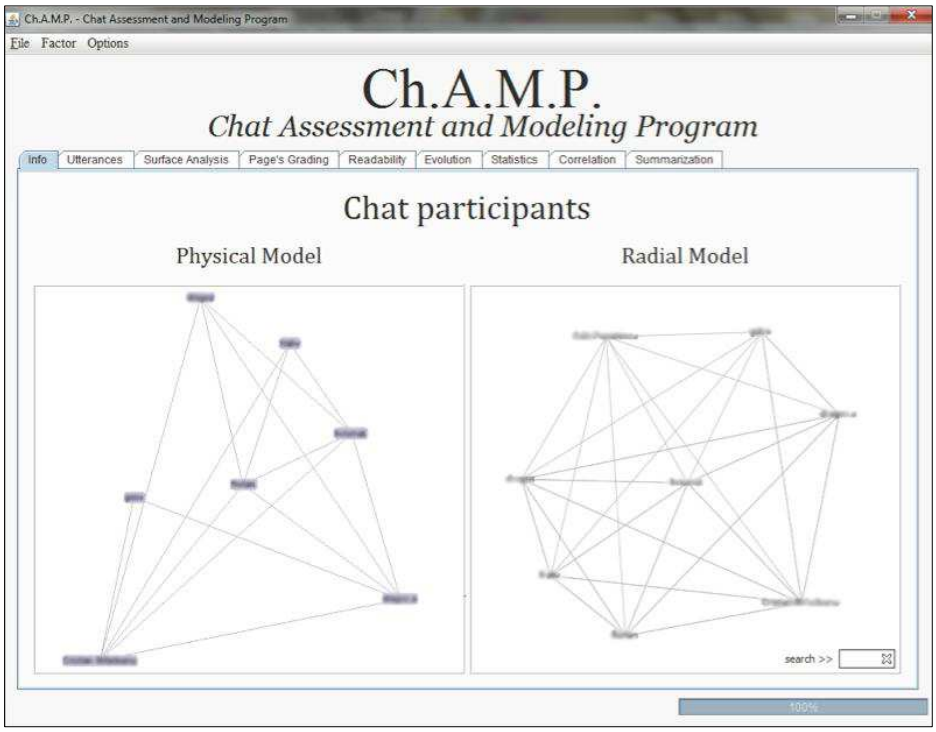


Figure 25. *Ch.A.M.P.* Main user interface.

The next step is grammatical and morphological analysis based on spellchecking, stemming, tokenization and part-of-speech tagging (Manning & Schütze, 1999). Eventually, a semantic evaluation is performed using Latent Semantic Analysis (Landauer & Dumais, 1997) for assessing the on-topic relevance score of each utterance based on a given set of keywords, predefined by the tutor to be covered throughout the conversation (more specifically, concepts of interest that need to be addressed – CSCL technologies).

Table 12. *Ch.A.M.P.* Evaluation hierarchy.

	Evaluation Hierarchy		Interaction modeling
1.	*Surface analysis*		
	Readability formulas and metrics derived from Page's essay grading techniques	–	Social Network Analysis
2.	*Morphological analysis* and *POS tagging*		
3.	*Semantic evaluation*	–	

Moreover, at surface and semantic levels, metrics specific to social network analysis (Freeman, 1977; Brandes, 2001; Nguyen & Hong, 2006) are applied for properly assessing the participants' involvement both from a quantitative (surface analysis) and qualitative (semantic evaluation) point of view (see Table 12 in which levels 1 and 3 have their sub-sequent analysis factors included).

A Utterance Scoring

After building the utterance graph that highlights the correlations between utterances on the basis of explicit references (made by the chat participants during the conversation) and after finishing the lexical, morphological and semantic analysis of each intervention, a new scoring mechanism consisting of three steps was proposed (Dascalu, Trausan-Matu, et al., 2010b).

1. *Evaluate each utterance individually* by taking into consideration the following features (besides the ones already considered in *A.S.A.P.*):

- The branching factor corresponding with the actual number of derived utterances from current one;
- The similarity of the current utterance with the overall chat;
- The overlap and the similarity of the current utterance with the given set of keywords assigned by the tutor.

Furthermore, this individual utterance scoring mechanism combines the *quantitative approach* (e.g., the length of the sentence based on the assumption that a piece of information should be more valuable if transmitted in multiple messages, linked together, and expressed in more words, not only to impress others, but also to be meaningful in the given context) with a *qualitative one* (the use of LSA and of predefined keywords).

2. *Emphasize the utterance score*

The assumption for this step was that the normal evolution of a conversation consisted of the following key moments: 1/ introduction, 2/ statement of interest, 3/ concrete or on-topic exchange of utterances, 4/ conclusions and 5/ final salutations, with a distribution of importance similar to a Gaussian distribution centered on the third moment. Therefore, for each conversation thread obtained by chaining utterances based upon the explicit links, the utterance individual scores were

increased correspondingly with a Gaussian distribution centered on the global maximum value established from the first step of the utterance scoring process and having a spread equal to the thread's length.

3. *Determine the final score* for each utterance *in the current thread*

Based upon the previous augmented value, the final score of each utterance is computed in terms of the previously connected utterance from the same discussion thread (see Equation 6):

$$score_{final}(u) = score_{final}(v, v \leftarrow u) + coefficient * score_{individual}(u) \quad (6)$$

where $score_{individual}(u)$ are the values obtained from the previous step, $score_{final}(v, v \leftarrow u)$ expresses the cumulative score per discussion thread taking into consideration the explicit links $v \leftarrow u$ and a $coefficient$ that is extracted from a predefined matrix based on the identified speech acts (Trausan-Matu et al., 2004). These predefined values were determined after analyzing and estimating the impact of the current utterance by considering only the previous one from the discussion thread (similar to a Markov process). As coefficients can be also negative when identifying negation speech acts, the cumulative score ($score_{final}$) of a discussion thread may be raised or lowered by each utterance. Therefore, depending on the type of an utterance and the identified speech acts, the final score might have a positive or negative value. The tendency was to observe whether we were dealing with a constructive conversation or with a lot of disagreement that would be represented as a negative slope in terms of the evolution of $score_{final}$.

Nevertheless, the context for introducing these steps is important as it took into consideration two dimensions: 1/ providing an overall view of the evolution of each discussion thread and the conversation as a whole (although, at this step the utterance graph was expressed only through explicit links) and 2/ reducing the negative impact of exceptional behaviors (e.g., if a user tries to impress others in terms of the complexity of his/her interventions, but without making much sense, therefore artificially increase his/her score, this component reduces this effect as those interventions would be isolated in terms of the main discussion threads and their corresponding importance will be diminished).

B Social Network Analysis applied on the Interaction Graph

Social factors are considered as an addition to the quantitative and qualitative measures computed at utterance level. From the point of view of social network analysis, various metrics are computed in order to determine the most competitive participant in chat: degree (in-degree, out-degree), centrality (closeness centrality, graph centrality and eigenvalues) (Brandes, 2001; Liu, 2011) and user ranking similar to the well-known Google Page Rank Algorithm (L. Page, 2001). All these SNA factors are applied on two interaction graphs. One is built by considering the effective number of interchanged utterances between participants (a *quantitative* approach), while the other uses instead the sum of utterance scores for all discussion threads and provides the basis for a *qualitative* evaluation.

All the identified metrics used in Social Network Analysis are relative, in the sense that they provide scores relevant only compared with other participants from the same chat, not with those from other conversations. This is the main reason for scaling factors between all participants by assigning each conversation participant a weighted percentage from the overall performance. In the end, the final score is a linear combination of all evaluation factors including surface analysis factors and SNA metrics applied on both interaction graphs, with their corresponding weights.

C Optimizing each Metric's Weight in the Final Participant Score

The goal of the designed algorithm is to determine the optimal weights for each given factor in order to have the highest correlation with the manual annotator grades. Pearson correlation was used in conjunction with the average error for better grasping a global tendency of the evaluation process; in the end, correlation was considered the determinant factor of evaluation as it is more cognitively and statistically relevant and less prone to bias. Moreover, a series of constraints had to be enforced for properly defining the weight optimization algorithm:

- (Optional) A minimum/maximum value for each weight – for example, a minimum of 2% for imposing the mandatory consideration of each evaluation factor and a maximum of 40%, allowing all factors to play a role in the final score and restricting the dominance of a single factor;
- (Hard constraint) The sum of all factors must be 100%;
- (Goal) Obtain a maximum mean correlation for all chat conversations.

In this case, the system integrated two components (Dascalu, Trausan-Matu, et al., 2010b):

- A *perceptron* (Rosenblatt, 1957; Collins, 2002) used for obtaining fast solutions as inputs for the genetic algorithm. The main advantages for adding this simple neural network reside in its fast convergence, numerical stability and the rapid search in the weight space for the (sub-)optimal solution;
- A *genetic algorithm* (Whitley, 1994; M. Mitchell, 1996; T. Mitchell, 1997) used for determining the optimal weights of each evaluation factor; this step can be considered a fine-tuning of the solutions given by the perceptron, constrained by the previously defined clauses and converging towards a local optimal solution.

The later algorithm operates over populations of candidate solutions or chromosomes that model a set of properties, more specifically weights for each evaluation factor taken into consideration in *Ch.A.M.P.* (Whitley, 1994). A population advances iteratively and each generation represents and approximation of the solution; therefore each subsequent generation is a refinement towards the determination of optimal weights in order to assure the best overall correlation. Correlation is expressed as an arithmetic mean of all correlations per chat conversations because of different evaluation styles; a concatenation of all scores would not be relevant because the system grades participants relatively to the best one from that specific conversation (local importance) and there was a high fluctuation in the grade distributions/spreads between different evaluators.

In the end, the genetic algorithm's goal was to maximize the overall correlation and, for maximizing the chances of finding an appropriate solution, specific adjustments have been made in terms of the proposed workflow (Dascalu, Trausan-Matu, et al., 2010b) (see Figure 26):

- A fixed number of 100 chromosomes per population seemed most appropriate (more than 6 times the number of cumulative evaluation factors).
- *Initialization*: 2/3 of the initial population is initialized outputs of the perceptron, while the rest is randomly generated in order to avoid only local convergence.
- *Fitness* function – the mean overall correlation for all chats in the corpus.
- *Selection* function – roulette based or elitist selection: the higher the fitness, the greater the possibility a chromosome is selected for crossover.
- *Crossover* function – based on real intermediate recombination (Muhlenbein & Schlierkamp-Voosen, 1993) which has the highest dispersion of newly generated

weights – select a random alpha for each factor between [-0,25; 1,25]; the relative distance between 2 chromosomes selected for crossover must be at least 20% in order to apply the operator over them.

- *Mutation* function – each weight had a probability p of being mutated at the current step by adding to or decreasing its value with a random number within a predefined range (1/2 of the variable domain) multiplied by a delta factor equal to $\sum_i a(i)2^{-i}$, where i is the index of the factor in the chromosome and $a(i)$ is 1 with p probability and 0, otherwise.
- *Correction* function – a necessary operator in order to assure that the initial constraints are satisfied: if values are above or below minimum/maximum, weights are reinitialized starting from the threshold and adding/subtracting a random quantity from it; if overall sum of percentages is different from 100%, adjust randomly weights with steps of 1 divided by the imposed precision.

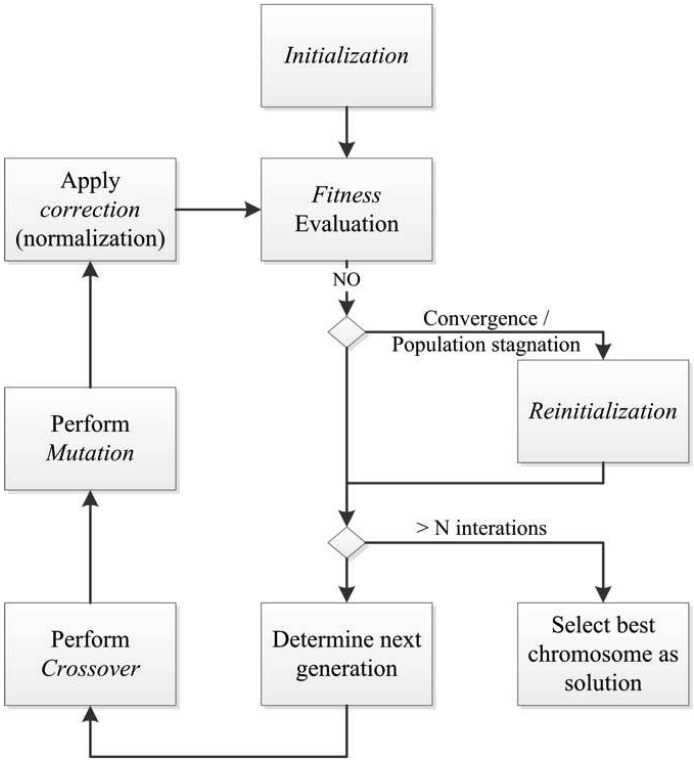


Figure 26. *Ch.A.M.P.* Genetic algorithm workflow.

In addition, three optimizations were integrated in the proposed genetic algorithm (Dascalu, Trausan-Matu, et al., 2010b). Firstly, we used a *CHC optimization* (Eshelman, 1991), with a slight modification – N children were generated and 20% of the best newly generated chromosomes were retained; 20% of the best parents were also kept in the new generation, whereas the rest of the

population consists of the best remaining individuals. Secondly *multiple populations* were concurrently generated (4 in most of the experiments) that exchanged their best individuals. Therefore, after 10 generations each population added its best individual to a common list and replaced its worst individual with a randomly selected one from the list. Thirdly, after a population has stagnated or reached convergence seen as having consequently the same best individual for 20% of the maximum number of generations, it was partially *reinitialized*: 10% of new population used the best individuals from the previous one, 30% was obtained as outputs of the perceptron and the remaining individuals were randomly generated.

In the end, the proposed solution for determining the optimal weights combined the two approaches in order to obtain benefits from both – numerical stable solutions from neural networks and the flexibility of genetic algorithms in adjusting these partial solutions.

5.2.3 Validation of *Ch.A.M.P.*

The initial running configuration for *Ch.A.M.P.* used the following predefined weights that were chosen to augment the semantic dimension of the analysis (overall, 60% was attributed to semantically related factors and only 40% to surface and quantitative SNA factors): 10% for the mixture of surface factors, 5% for each social networks factor applied on the number of interchanged utterances and 10% for each semantic social network metric built on the interaction graph that integrates the utterance scores.

The validation was performed on 23 conversations (4th year computer science students debating on CSCL technologies in groups of 4-5 members) and the preliminary results, without any optimizations and by using only the previous weights, were: average error of 3 on the [0; 10] grading scale and r = .514. The average error was lower than *A.S.A.P.* as the experiment was conducted on a completely different corpus, in which the peer grades were reviewed and adjusted by a tutor; therefore in the end, although two iterations were performed, only one score was assigned per participant – that of the tutor – as we were dealing with an extremely time consuming process. Moreover, as the number of evaluation factors increased, the weight of the quantitative ones, also used in *A.S.A.P.*, decreased correspondingly.

After running the weight optimization algorithm, the synthetic results in terms of the factors with the highest weights, after multiple runs with 4 concurrent populations (see Figure 27), are presented

in Table 13. Due to the fact that the first three factors that emerged as being the dominant ones for best predicting the automatic scores were based only on rather simple and straightforward quantitative factors, we can clearly observe a predominantly quantitative approach in the human grading process. All remaining factors were evaluated below 5% and, therefore, did not have a high importance in the final grading process.

Table 13. *Ch.A.M.P.* Evaluation factors with an importance greater than 10% after multiple runs of the weight optimization algorithm.

Percentage	Factor
30-40%	Out-degree expressed by the number of outgoing utterances – somehow a participant's gregariousness measure
20-25%	Aggregated surface analysis factors extracted from Page's initial studies
10-15%	In-degree applied on number of interchanged utterances
≈ 10%	Semantic graph centrality – the only measure qualitative measure with a higher importance that relies on utterance scores

The overall results, with regards to correlation optimization, obtained after running the genetic algorithm were: average error of 5.4 and r = .594 as the average correlation between the human grades and the automatic scores for all conversations. The increase of the average error was quite natural, as the automatic scoring mechanism grades each conversation independently; therefore, each time at least a participant receives 10, although this is inappropriate for some conversations.

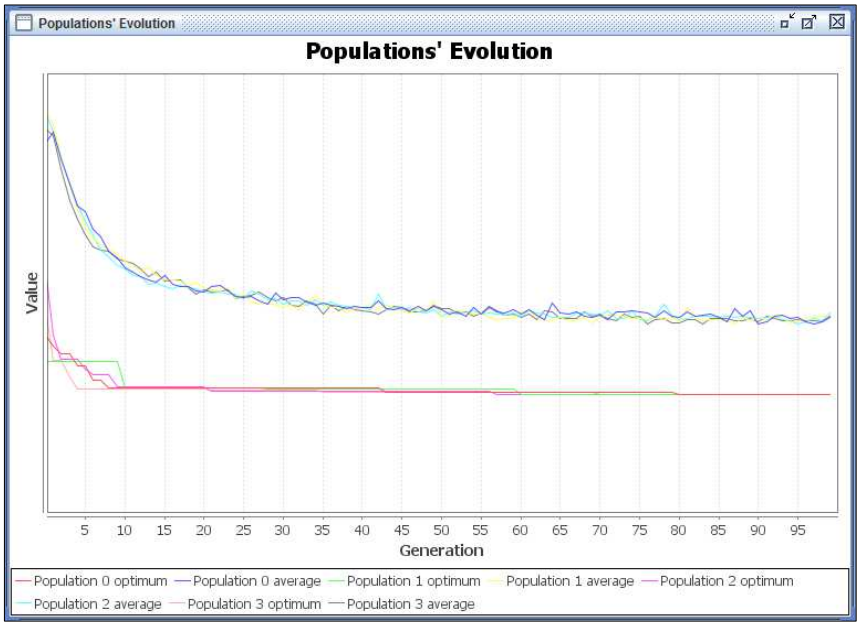


Figure 27. *Ch.A.M.P.* Convergence to an optimal solution using 4 concurrent populations, with the visualization of optimum/average fit of chromosomes.

The spikes in Figure 27 from each population's average fitness are determined by newly inserted individuals or by the population reinitialization. After the first 10 iterations, important improvements can be observed, whereas after 30 generations the optimum chromosomes of each population stagnate and we can consider that convergence was achieved. Only population reinitializations and chromosome interchanges provide minor improvements.

As conclusion, all the previous results obtained from *A.S.A.P.* and *Ch.A.M.P.*, as well as the valuable technical expertise gathered while developing the actual systems, were taken into consideration in the implementation of subsequent systems. Moreover, while looking at *A.S.A.P.* and *Ch.A.M.P.*, a multiple key aspects were identified. Firstly, the final scoring of participants and the general grade that was presented to the student had no actual impact on him/her, as it did not reflect his/her strong or weak points through the conversations. In the end, it was a mere number assigned after performing a post-conversation analysis of the chat log whose significance was hard to grasp. Moreover, the multitude of factors that were presented within the interface and were integrated in the final score had no significance to the learner, as no corresponding interpretations were provided in order to better grasp the cause-effect relations that might help a student improve in subsequent conversations.

Secondly, besides involvement that was clearly encouraged, there was no emphasis on collaboration or on the effect of social knowledge-building. One could talk extensively, using an elevated vocabulary, following the predefined tutor list of concepts, but without any interaction with other participants. Although the scoring mechanism tried to estimate the importance of each intervention, it could not grasp this collaborative dimension, thus significantly diminishing the overall results of the learning activity.

Thirdly, the conducted assessments were centered only on the correctness of the final scores, not on their interpretation or later usefulness. Moreover, only an overall grading perspective in terms of involvement was followed, without an emphasis on collaboration. In addition, only a local perspective, over each individual was taken into consideration. Nevertheless, a global perspective of the community should have been used instead of the average local correlations. Participants should be evaluated in comparison one to another, but across multiple conversations, not based on a local scaling. In the end, as newer discourse structures, like the cohesion graph (Trausan-Matu, Dascalu, & Dessus, 2012), have more inter-dependencies and a more dense structure, the second and the third steps proposed in the *Ch.A.M.P.* utterance scoring mechanism proved improper as, in some

cases, a great span of cohesive utterance would have had their corresponding scores augmented in an unnatural manner, whereas multiple cohesive links already induce this augmentation effect.

Thus, a shift towards providing comprehensive feedback is required, as the metrics integrated in *A.S.A.P.* and *Ch.A.M.P.*, seen as individual factors, do not provide sufficient insight to learners and tutors. Moreover, in terms of CSCL conversations, collaboration should play a central role in the conducted analysis, even when considering the assessment of participants' involvement. All these identified effects and their educational implications were addressed in *PolyCAFe* and, later on, in *ReaderBench*, that are presented in detail in Chapters 6-9.

6 *PolyCAFe* – Polyphonic Conversation Analysis and Feedback

The *PolyCAFe* system (Polyphonic Conversation Analysis and Feedback generation) (see Table 14) was designed, implemented and validated within the FP7 2008-212578 LTfLL – Language Technologies for Lifelong Learning project (Trausan-Matu et al., 2008; Trausan-Matu, Dessus, et al., 2010; Trausan-Matu et al., 2011). Moreover, it represented the starting point of the joint work between the two universities (University Politehnica of Bucharest and University Grenoble Alpes), as both were partners in the same work package.

Table 14. *PolyCAFe* Traceability matrix of provided functionalities and integrated tools.

	Tools						
Functionality	Patterns	IR	NLP pipe	*WordNet*	LSA	SNA	Distributed computing
Utterance graph and conversation visualization	✓		✓		✓	✓	
Speech acts identification	✓	✓					
Topics modeling		✓			✓		
Dialogical perspective	✓				✓		
Participant involvement evaluation			✓		✓	✓	✓
Intervention scoring		✓	✓		✓	✓	✓
Collaboration assessment					✓		
Comprehensive feedback delivery			✓		✓	✓	
Semantic search		✓		✓	✓		
Semantic extractive summarization		✓			✓	✓	

6.1 General Presentation

As seen in *3.1.3* CSCL Computational Approaches, collaborative applications on the web were constantly developed in the last years in many domains and one remarkable case is their usage for educational purposes. In particular, Computer Supported Collaborative Learning (CSCL) (Stahl, 2006a, 2006b) is very well suited from both practical and theoretical reasons. As commodity, chats and forums are commonly used by students as they offer the possibility of joint learning anytime and anywhere. Thus, CSCL became a viable alternative or supplement to classical learning when targeting small virtual groups using chat systems for learning together (Koschmann, 1999; Stahl, 2009b). Moreover, a paradigm shift occurred in the sense that learning can be achieved through the participation to a dialogue that constructs discourse, rather than a transfer of knowledge from teachers or textual documentation to students (Bereiter, 2002; Stahl, 2006b; Trausan-Matu et al., 2006). In essence, the CSCL paradigm is based on dialogism and the social-cultural ideas of Vygotsky (Vygotsky, 1978; Cazden, 1993) and Bakhtin (Bakhtin, 1981, 1984) that appeared decades before the invention of the computer (as clarification, the original work appeared long before the cited translations, in the first decades of the twentieth century).

As the automatic analysis of conversations is a difficult task, *PolyCAFe* combines approaches from previous systems focused on chat analysis (Trausan-Matu, Rebedea, et al., 2007; Dascalu et al., 2008a; Dascalu, Rebedea, et al., 2010; Dascalu, Trausan-Matu, et al., 2010b) and provides abstraction and feedback services for supporting both learners and tutors involved in assignments that make use of chat or forum conversations. In order to respond to the previous challenge of providing relevant feedback after a thorough analysis of the conversation, *PolyCAFe* integrates the *dialogistic polyphony* model (Trausan-Matu, Stahl, et al., 2007; Trausan-Matu & Rebedea, 2010) (see *3.1.2* Bakhtin's Dialogism as a Framework for CSCL and *4.2* Discourse Analysis and the Polyphonic Model) with *social network analysis* (Dascalu, Trausan-Matu, et al., 2010b) (see *3.2* Social Network Analysis), *information retrieval* (Adams & Martell, 2008; Manning et al., 2008), *machine learning* (T. Mitchell, 1997) and *natural language processing* (Manning & Schütze, 1999; Jurafsky & Martin, 2009; Dascalu, Trausan-Matu, et al., 2010b; Trausan-Matu & Rebedea, 2010) (see *4.3* Natural Language Processing Techniques). It was developed as one of the modules of the Language Technologies for Lifelong Learning project (LTfLL, see http://www.ltfll-project.org) (Berlanga, Van Rosmalen, Trausan-Matu, Monachesi, & Burek, 2009) funded by the European Commission under

the 7th Framework Programme. All provided services are packed into web widgets that can be easily integrated into any LMS, PLE, VLE or other web applications (e.g. blogs that use Wordpress) and the online version of the system is available at the following address: http://ltfll-lin.code.ro/ltfll/wp5/.

6.2 Theoretical Considerations and Educational Scenario

As mentioned before, *PolyCAFe* and its precursor *Polyphony* (Trausan-Matu, Rebedea, et al., 2007) are probably the first systems that implement Bakhtin's ideas on dialogism (Bakhtin, 1984; Koschmann, 1999; Stahl, 2006b), with emphasis on inter-animation and polyphony (Bakhtin, 1984; Sarmiento et al., 2005; Trausan-Matu et al., 2005; Trausan-Matu, Stahl, et al., 2007; Dessus & Trausan-Matu, 2010) (see *3.1.2* Bakhtin's Dialogism as a Framework for CSCL). The importance of considering these ideas is that they allow understanding the mechanisms of collaboration and they provide a theory that can be used to measure the contributions of participants in chats and forums (Trausan-Matu & Rebedea, 2009, 2010).

There are several ways in which Bakhtin's ideas are used in the implementation. First of all, several voices are considered to be present and the aim consists of identifying them throughout each conversation. The concept of 'voice' has an extended range as it is considered a position, a thread, not only a sound emitter (Trausan-Matu, Stahl, et al., 2007; Trausan-Matu & Rebedea, 2009). The contribution of each participant is computed considering the "strength" of his/her involvement and also the degree of inter-animation, which is the degree in which voices overlap and refer to each other. The computation of the strength of voices and of inter-animation uses NLP techniques (Manning & Schütze, 1999), social network analysis (SNA) (Brandes, 2001; Mislove et al., 2007) and Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997; Landauer, Foltz, et al., 1998). LSA plays a key role for determining the similarity between concepts, for linking different voices and for adding a semantic dimension to our analysis, while SNA enriches this perspective by enabling a deep insight into personal involvement and evolution. Based on these premises, overall collaboration between chat participants is also automatically assessed (Dascalu, Rebedea, et al., 2010; Dascalu, Rebedea, et al., 2011).

There are many advantages (Eastman & Swift, 2002; Stahl, 2009b) for using chats in contexts that would involve collaborative problem solving, engaging in debates or stimulating the creativity of learners through brainstorming-like sessions. However, taking into consideration the difficulty and

the required time for providing feedback to students involved in such conversations (Trausan-Matu, 2010a), this scenario may become less appealing to teachers and decision makers in universities and schools.

PolyCAFe has been designed starting from the experience of participating as tutor/professor in using instant messenger (chat) for CSCL in two different settings. The first one is the Virtual Math Teams (VMT) project (Stahl, 2009a). The second is the usage of CSCL chats in a Human–Computer Interaction course for undergraduate senior year students, as well as for master students studying Adaptive and Collaborative Systems, Natural Language Processing, and Symbolic and Statistical Learning courses at University Politehnica of Bucharest. At these courses, students were given between 1–3 assignments that needed to be solved using a chat conversation in unmoderated small groups of around four participants (for a detailed presentation check sections *6.5.1* First Validation and *6.5.2* Second Validation). For example, in a typical assignment students were told to debate and argue for the best web communication and collaboration technology to be used by a company in a certain context (see Appendix D – Input Examples, Sample Chat – Log of Team 4 Chat Conversation for an excerpt highlighting different phases of a conversation). At another course, they had to discuss about the topic of the next lecture in order to identify and agree on the most important aspects and make a short collaborative slideshow about that topic. They had to read in advance relevant online and offline documents to be able to have a good understanding of the subject during the discussion.

After the students finished a chat conversation, the tutors read the transcript and provided a feedback and grading to the students. A major problem was that this proved to be an extremely difficult task, especially when the number of teams to analyze is large. As highlighted by Trausan-Matu (2010a), the manual analysis of chat conversations is extremely time consuming, lasting more (even two times) than the actual time spent by the participants during their online session. Therefore, a need of a computer system to help, facilitate and support teachers by greatly reducing the time spent and by providing a wide variety of metrics to gaze upon was identified.

Moreover, because in our experiments the feedback delivered by different tutors was not very consistent, a feedback schema for this type of assignments has been developed in *PolyCAFe* that defines the most important elements that should be assessed. Thus, it was decided that the feedback should be delivered on three distinct levels: for the conversation as a whole, for each participant and

for each utterance (or at least for the most important utterances in the conversation). Each of these types of feedback should take into account the content or domain knowledge, the collaboration, the involvement and the social impact of each participant.

PolyCAFe is a web–based system designed to be easily used by learners and tutors working in a similar educational scenario. It was designed starting from the polyphony inter–animation model (Trausan-Matu, 2010c). It implements Natural Language Processing (NLP) and Social Network Analysis (SNA) techniques and takes advantage of the experience acquired with the development of the previous systems. The system was designed with the idea that it will not eliminate the tutor's presence in the evaluation process, but only to provide him learning analytics support.

PolyCAFe provides feedback and support services based on the practice currently used by the tutors for assessing a chat conversation. This feedback is delivered to students and tutors after they finish a discussion in order to provide the final evaluation to the learners. In this manner, students get preliminary results from the system in order to understand what aspects need improvement (e.g., active involvement, language, on–topic relatedness), while the tutors are helped to provide a more qualitative, consistent evaluation, in less time.

6.3 Widgets Overview

In order to provide learners and tutors extensive control over the generated feedback interfaces, *PolyCAFe* was implemented as an online platform whose results are displayed in web widgets that can be used independently or together (Rebedea et al., 2010; Trausan-Matu, Dessus, et al., 2010; Rebedea, Dascalu, Trausan-Matu, Armitt, et al., 2011; Trausan-Matu et al., 2011). In this manner, the processing is decoupled from the interface and the widgets can be easily integrated into most online learning environments and other web platforms. *PolyCAFe* provides 2 management widgets and 5 feedback widgets. The *assignment management widget* enables tutors to define, edit and delete assignments. The *conversation management* widget enables create, read, update, and delete operations for conversations (chats or discussion threads from online forums).

One of the most important widgets is the *conversation visualization widget,* which displays a diagram of the utterances emitted by each participant (the "utterance graph"), represented as small rectangles aligned to the right of each participant's name, following the conversation timeline. The links between them are differently colored. For example, in Figure 28 explicitly mentioned links through

the referencing facility of the *VMT* or *ConcertChat* environments (Holmer et al., 2006) are marked as green, whereas the implicit links detected with NLP techniques are red. Moreover, by considering these links, discussion threads can be automatically identified. Users can zoom in on both vertical and horizontal axes of the utterance graph visualization (conversation participants vs. utterances spanning in time), enabling a detailed view of a conversation segment or an overview of the entire discussion (see the different timescales between the print-screens depicted in Figure 28). A graphical representation indicating a computed degree of collaboration is also presented in the same interface as a blue-colored graph depicted below the conversation graph (see Figure 28), concomitantly following the conversation timeline. A detailed presentation of how the collaboration degree is actually computed is included in *6.4.2* Collaboration Assessment.

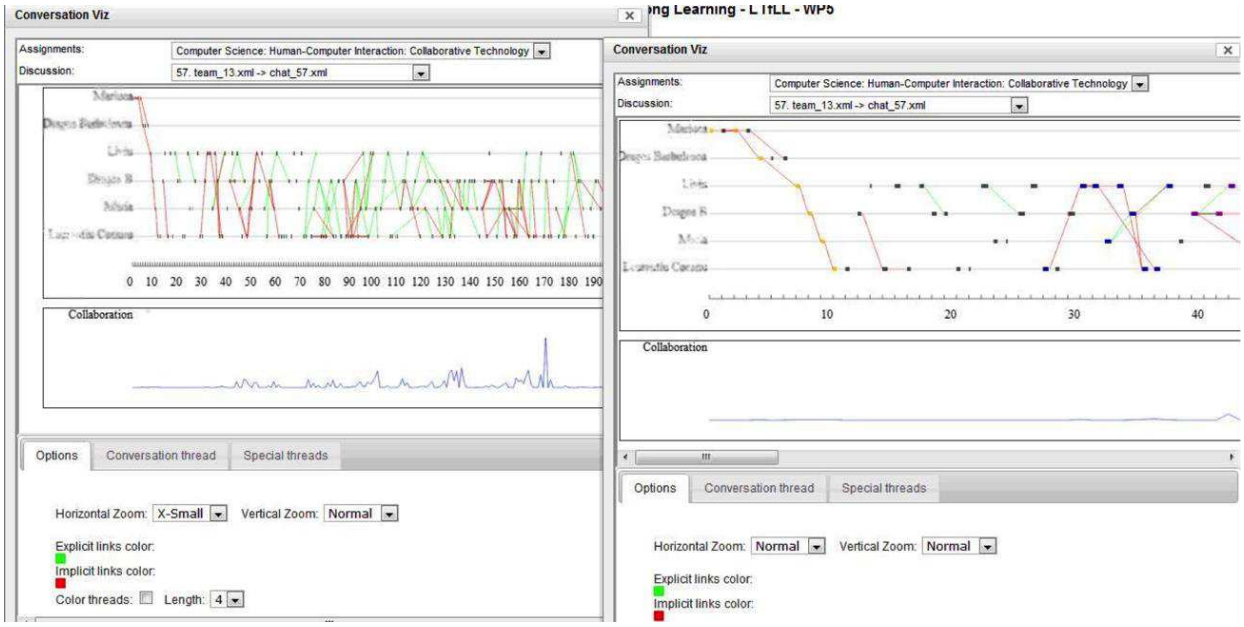


Figure 28. *PolyCAFe* Conversation visualization widget.

The *participant feedback widget* offers an assessment for each participant on several levels: relevance with regards to the domain corpora, social presence, importance or surface analysis factors (see Figure 29).

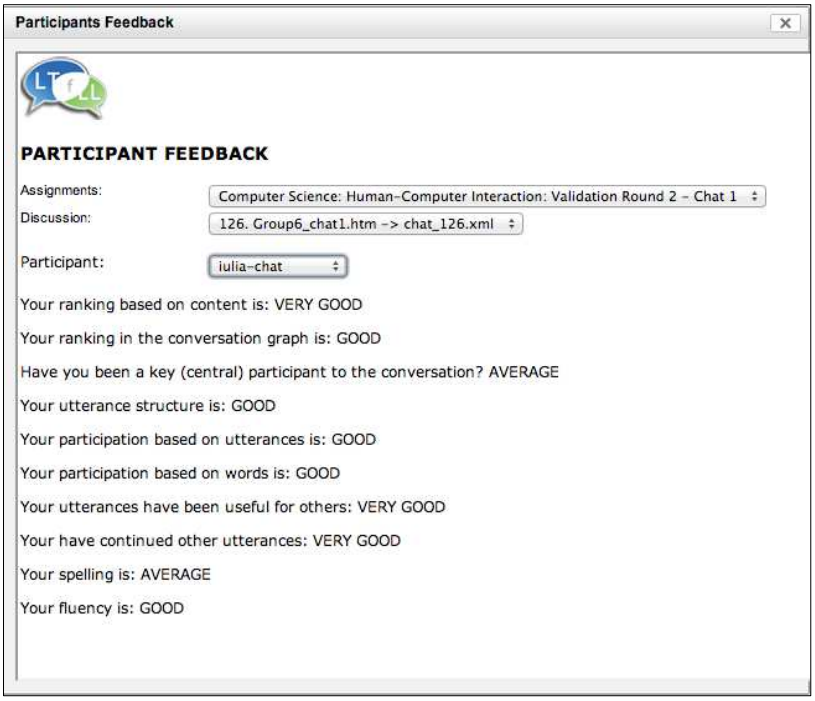


Figure 29. *PolyCAFe* Participant feedback widget.

The *conversation feedback widget* presents general information and statistics about the entire conversation: the most relevant concepts from the conversation, a suggestion of concepts from the semantic space that are semantically similar to the ones discussed in the chat and statistics regarding the density of the utterance graph or the percent of several types of dialog acts, such as personal opinions, request for information and arguments (see Figure 30).

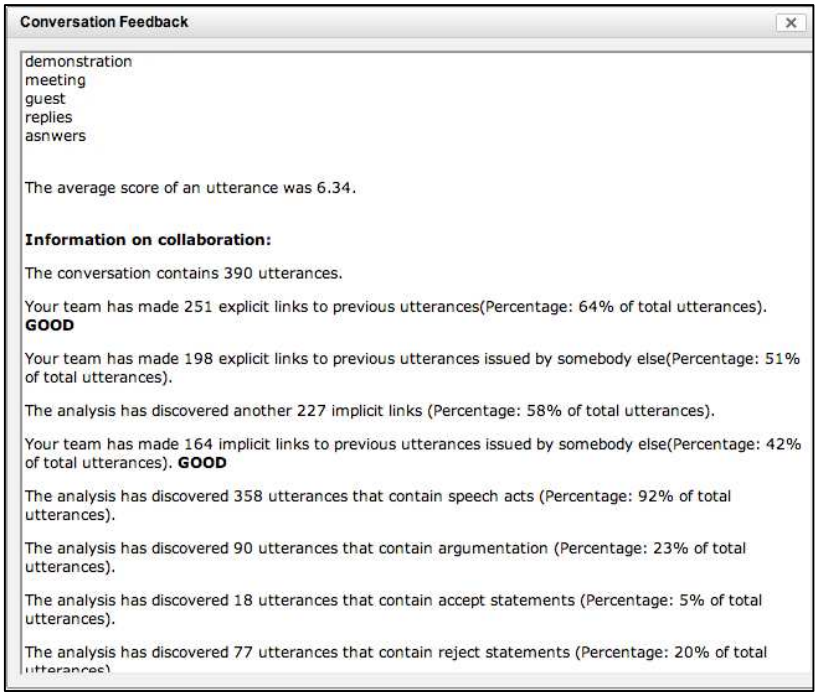


Figure 30. *PolyCAFe* Conversation feedback widget.

The *utterance feedback widget* gives indicators for each post in the conversation: speech acts and argumentation patterns that are present in the utterance, supplemented by a numerical value computed mainly by taking into account the lexical and semantic information of the text in the message (see Figure 34). Moreover, this widget also may present the users a summary of the conversation that includes only the most important utterances in the discussion with regard both to the content and to the collaborative discourse.

The *search conversation widget* provides a mechanism for ranking utterances and participants with regards to a search query provided by the user. It takes into consideration not just the lexical items, but also the semantic relations and the importance of each utterance as considered by the utterance evaluation process (see Figure 35).

6.4 Architecture and Core Functionalities

Technically, *PolyCAFe* consists of a series of NLP and SNA computations (Rebedea et al., 2010; Trausan-Matu, Dessus, et al., 2010; Trausan-Matu & Rebedea, 2010; Rebedea, Dascalu, Trausan-Matu, Armitt, et al., 2011) and its main tasks are implicit link detection using patterns, repetitions and semantic distances based on *WordNet* and LSA (Trausan-Matu & Rebedea, 2010; Rebedea, 2012), utterance evaluation and collaboration analysis based on the utterance graph. The result of these computations provides feedback on several distinct levels: for each utterance in the conversation, for each participant and for the conversation as a whole.

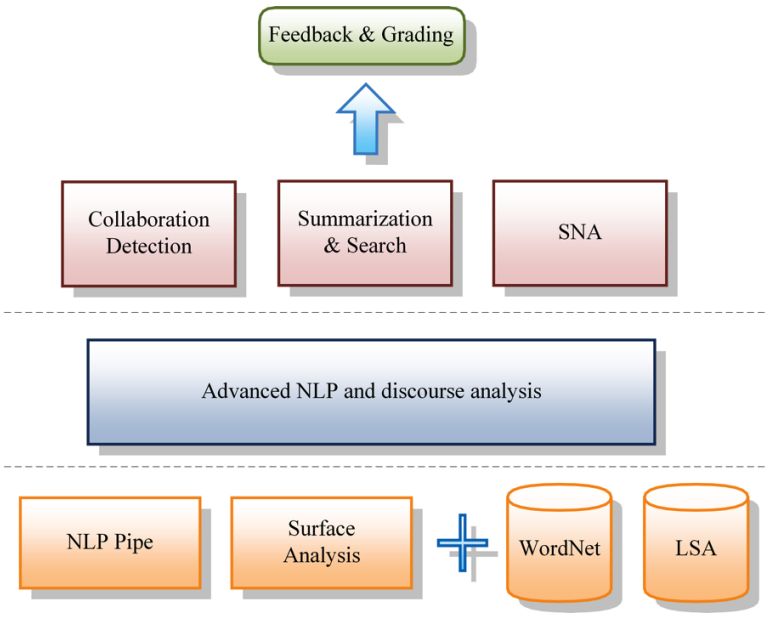


Figure 31. *PolyCAFe* Simplified technical architecture.

The *PolyCAFe* system has a multi–layered architecture as depicted in Figure 31. A more detailed diagram has been presented in (Rebedea, Dascalu, Trausan-Matu, Armitt, et al., 2011). The *first layer* contains basic NLP processing, surface textual evaluation, and concepts' extraction. The first step in analyzing the raw data is a typical NLP series of basic processing (a "NLP pipe"): spelling correction, stemming, tokenization, part of speech tagging and parsing (Manning & Schütze, 1999). Next, a surface analysis is performed consisting of computing metrics derived from Page's essay grading techniques (E. Page, 1966) and readability measures (Davison & Kantor, 1982; Brown, 1998). The *WordNet* (http://www.wordnet.edu) lexical database and LSA (Landauer & Dumais, 1997) semantic spaces compose the semantic resources sub–layer used for concept extraction. These two resources form also the basis for a semantic evaluation of the participants' involvement and evolution. In contrast with surface analysis based solely on quantitative measurements, WordNet and LSA enable a qualitative assessment of the overall discussion and of the involved participants.

The *second layer* contains advanced NLP and discourse analysis modules for the automatic identification of underlying interactions among participants (Dascalu, Trausan-Matu, et al., 2010b; Trausan-Matu & Rebedea, 2010). To this aim, speech acts, lexical chains, adjacency pairs, co–references and semantic similarities are identified. All these are the starting points for candidates of implicit links that constitutes the arcs in the *utterance graph* (Trausan-Matu & Rebedea, 2009; Dascalu, Trausan-Matu, et al., 2010b), in addition to the explicit links indicated by participants as references to previous utterances in the chat environment room (Holmer et al., 2006; Stahl, 2009a). This structure plays a central role in the scoring process of each utterance and of each participant (Dascalu, Rebedea, et al., 2010). In addition, threads are identified using specific graph algorithms (Cormen et al., 2009), one of the most simple ones being the identification of the connected components in the conversation graph, while other methods using graph flow could also be employed.

The *third layer* takes into account the detection of collaboration (or collaborative discourse) starting from the analysis of the utterance graph with SNA techniques and following the polyphony analysis method (Trausan-Matu et al., 2005; Trausan-Matu, Stahl, et al., 2007). SNA specific metrics are also computed on the utterance graph for identifying the central utterances within each discussion thread (Dascalu, Rebedea, et al., 2010; Dascalu, Rebedea, et al., 2011). Therefore, SNA is used at a surface level for modeling the interaction between participants as number of interchanged utterances, but also at a deep, semantic layer by taking into consideration the score of each utterance determined

by means of LSA (Dascalu, Rebedea, et al., 2010). Moreover, the individual involvement of participants derived from SNA must be correlated with collaboration assessment, as a conversation becomes more interesting as its underlying discourse is non-linear and as its utterances are intertwined in a polyphonic manner (Trausan-Matu et al., 2005).

The final step in the analysis consists of aggregating all the factors obtained as outputs from the previous sub–layers and displaying them in an intuitive manner within the user interface, in order to offer textual and graphical feedback and an assessment proposal for each participant of the chat conversation. Extractive summarization and semantic search are easily achievable through the interaction with the previous components and by making use of the conversation graph, therefore they are provided as additional features of *PolyCAFe*. In addition, a distributed computing architecture (Dascalu, Dobre, et al., 2011) was enforced in order to enable parallel analysis of corpora containing multiple conversations.

This section continues with a computational perspective on the representative tasks of *PolyCAFe:* the process of evaluating the importance of an utterance, the assessment of collaboration within the conversation, extractive summarization, semantic search and the distributed architecture.

6.4.1 Utterance Evaluation

In order to obtain a thorough evaluation of involvement and of collaboration, the first step that needs to be undertaken is the actual evaluation of the utterances' importance within the conversation. This process involves building the utterance graph that highlights the intertwining of utterances, the determination of each utterance's importance in a given context and the assessment of on–topic relevance relative to the entire discussion.

The scoring process of each utterance is based on the utterance graph in which the weights of the arcs are computed as a semantic similarity function between the utterances, using LSA. Starting from the explicit links added by the user within the conversation environment and from the automatically identified implicit links, the graph is built by transposing these arcs, therefore following the timeline of the chat and the evolution of the current discussion in time. The actual scoring process of each utterance has three distinctive components: a quantitative, a qualitative and a social one that are briefly detailed in Table 15 (Dascalu, Rebedea, et al., 2010; Dascalu, Rebedea, et al., 2011).

Table 15. *PolyCAFe* Utterance evaluation hierarchy.

Component	Factor
Quantitative	NLP Pipe for utterance processing (spell-checking, stemming, tokenizing, POS tagging)
	Number of characters for each stem and corresponding number of occurrences as base of evaluation
Qualitative	Semantic similarity based on LSA for assessment
	Predefined topics coverage
	Thread evolution with regards to future impact and thread cohesion
	Overall discourse impacting utterance relevance
Social	Social Network Analysis applied on the utterance graph (degree)

The *quantitative perspective* (see Equation 7) evaluates each utterance at surface level. Solely from this view, the assigned score considers the length in characters of each remaining word after eliminating those that do not carry content, for example, "a", "the", "to", etc., spellchecking and stemming (extracting the root of the words, by eliminating suffixes like "ed", "ing", "ly", etc.) are applied. In order to keep the inputs of the system as clean as possible, only dictionary words are considered. Moreover, as it is a common practice to use abbreviations in CSCL conversations, a list of translations is used to expand the shortened versions encountered in the discussion. In the end, we apply the logarithm function on the number of occurrences of each word (Manning et al., 2008) for reducing the impact of unnecessary repetitions used only for artificially enhancing the score of each intervention.

$$quantitative(\mathrm{u}) = log\left(\sum_{\substack{\text{remaining}\\ \text{words}}} \mathrm{length(stem)} \times \mathrm{Tf(stem)} \times \mathrm{Idf(stem)}\right) \quad (7)$$

The *qualitative dimension* involves the use of LSA in determining four different components: *thread cohesion*, *future impact*, *relevance* and *topics coverage*.

Starting from the utterance graph, *thread cohesion* (see Equation 8) for a given utterance represents the percentage of links, starting from that specific utterance, which share a similarity above a given threshold. Thus, thread cohesion is a backward-looking mechanism used for assessing the importance of an utterance within the ongoing discussion threads. In order to ensure inner cohesion and

continuity within a thread, similarities between any adjacent utterances must exceed the specified threshold (in our case 0.1).

$$thread\ cohesion(u) = \frac{no\ links_{u \to v} with\ sim(u,v) \geq threshold}{no\ links_{u \to v}} \quad (8)$$

Future impact (see Equation 9) enriches thread coherence by quantifying the actual impact of the current utterance with all inter-linked utterances from all discussion threads that include the specified utterance. It measures the information transfer from the current utterance to all future ones (explicitly or implicitly linked) by summing up all similarities above the previously defined threshold. In terms of the polyphonic model based on Bakhtin's dialogism (Bakhtin, 1981, 1984), future impact resembles echo, as it measures the information transfer from the current utterance to all future ones (explicitly or implicitly linked) by summing up all similarities above the previously defined threshold. Future impact, therefore the echo of a given voice, is estimated by measuring similarity between the two linked utterances.

$$future\ impact(u) = 1 + \sum_{link\ u \to v} sim(u,v) \quad (9)$$

Relevance (see Equation 10) expresses the importance of each utterance with regards to the overall discussion. This can be easily measured by computing the similarity between the current utterance and the vector assigned to the entire conversation, therefore determining the correlation with the overall discussion.

$$relevance(u) = sim(u, entire\ conversation) \quad (10)$$

Because each discussion has a predefined set of topics that had to be followed and which should represent the focus concepts of the chat, *topics coverage* (see Equation 11) measures the coverage of these keywords in each utterance. In our implementation, topics coverage is obtained by evaluating the similarity between each utterance and the specific set of keywords specified by the tutor or teacher as important topics of the discussions by means of semantic distances and cosine similarity within the LSA vector space. In other scenarios or for other tasks, the initial topics can be computed automatically from a given corpus that should be read by the students prior to the discussion.

$$topics\ coverage(u) = sim(u, predefined\ list\ of\ concepts) \quad (11)$$

The social dimension (see Equation 12) implies an evaluation from the perspective of social network analysis performed on the utterance graph. In the current implementation only two measures from graph theory (Cormen et al., 2009) are used (in-degree and out-degree), but other metrics specific to SNA (Freeman, 1977; Brandes, 2001; Newman, 2010) and minimal cuts (Cormen et al., 2009) will be considered.

$$social(u) = \prod_{\substack{all\ social\ factors\ f \\ quantative\ or \\ qualitative}} (1 + log(f(u))) \qquad (12)$$

In order to provide a clearer image of the previous metrics, Figure 32 presents a slice of a conversation that could also represent a partial discussion thread centered on the utterance that is under analysis, with the demarcation of possible links (explicit or implicit) from the utterance graph and with the presentation of the considered analysis factors.

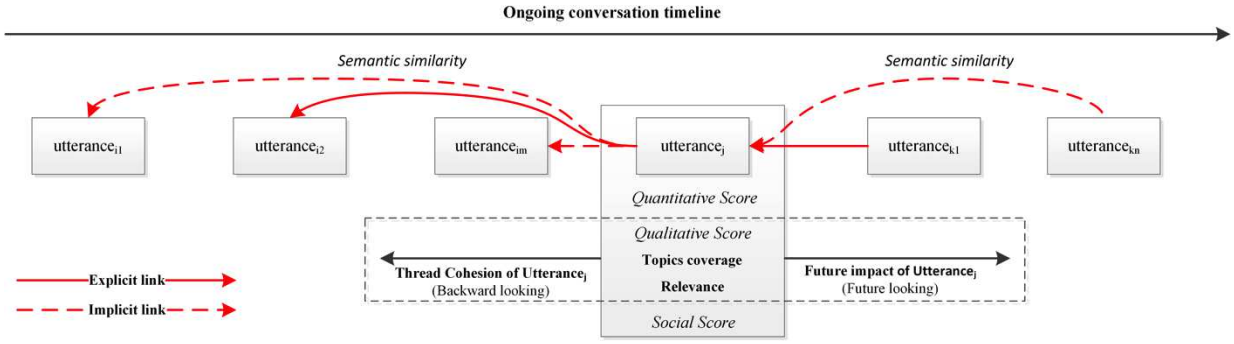


Figure 32. *PolyCAFe* Slice of the utterance graph emphasizing the utterance analysis factors.

6.4.2 Collaboration Assessment

Knowledge may be built in two different manners, each effecting the individual: *personal knowledge–building* (building personal knowing) (Stahl, 2006b) when new information is derived through self–study and self–experience and *collaborative learning*, through *social knowledge–building* by interacting with other people (Scardamalia, 2002). The concept of gain (Dascalu, Rebedea, et al., 2010) may be used for evaluating the contribution of each utterance in the overall discourse. It is derived from information theory (Shannon, 1948; Kent, 1983) and it is used for highlighting the importance and the future impact of the current utterance by taking into consideration all previous inter–linked utterances.

Starting from the two manners of knowledge-building, the following types of gain can be defined: *personal gain* when the interlinked utterances have the same speaker and *collaborative gain* when further information in the discussion thread is given by a different participant (Dascalu, Rebedea, et al., 2010). Moreover, the concept of gain (Dascalu, Rebedea, et al., 2010) is tightly related to echo. Individual echoes, which assume voice internalization and reflection as personal continuation of alien voices pertaining to different participants, can be transposed into personal gain, whereas collective echoes, responsible for context enrichment between different participants, can be assimilated to the concept of collaborative gain.

From the computational perspective, each utterance gets an importance score determined using the previously presented evaluation process (see Equation 13) and an overall gain composed of personal and collaborative gains (see Equations 14, 15 and 16). Furthermore, personal and collaborative gains are obtained by summing up the utterance importance score and the gain of the previous inter–linked utterances multiplied by the similarity between the previous interventions and the current one, depending on the type of interlocutors – same or different speakers (the equation for the utterance mark is introduced as recall to the previously proposed evaluation hierarchy – see Table 15).

$$mark(\mathrm{u}) = quantitative(\mathrm{u}) \times qualitative(\mathrm{u}) \times social(\mathrm{u}) \quad (13)$$

$$gain(\mathrm{u}) = personalGain(\mathrm{u}) + collabGain(\mathrm{u}) \quad (14)$$

$$personal\ gain(\mathrm{u}) = \sum_{\substack{\text{link u}\to\text{v} \\ \text{with } \mathbf{same} \text{ speaker}}} \big(mark(\mathrm{v}) + gain(\mathrm{v})\big) \times sim(\mathrm{u},\mathrm{v}) \quad (15)$$

$$collaborative\ gain(\mathrm{u}) = \sum_{\substack{\text{link u}\to\text{v} \\ \text{with } \mathbf{different} \text{ speakers}}} \big(mark(\mathrm{v}) + gain(\mathrm{v})\big) \times sim(\mathrm{u},\mathrm{v}) \quad (16)$$

Combining the utterance importance score with the gain gives us a good estimation of the actual importance of an utterance in a given context, while cosine similarity (Manning & Schütze, 1999) measures the strength, the impact and the echoes between the two explicitly or implicitly inter-linked utterances. By summing up all previous influences, we obtain a clear estimation of the retrospective effect for each utterance.

In the end, collaboration for the entire discussion is evaluated by comparing the overall collaborative gain of all utterances to the sum of all individual utterance marks or to the sum of overall gains. These two measures provide the means to determine the percentage of actual collaboration, not monologue, within the discussion: *mark-based collaboration* (see Equation 17) expresses the percentage of information that is built/transferred in a collaborative manner, whereas *gain-based collaboration* (Equation 18) weights the collaborative gain relative to the overall gain (Dascalu, Rebedea, et al., 2010).

$$\text{mark based collababoration} = \frac{\sum_{\text{all utterances u}} collaborative\ gain(\text{u})}{\sum_{\text{all utterances u}} mark(\text{u})} \quad (17)$$

$$\text{gain based collaboration} = \frac{\sum_{\text{all utterances u}} collaborative\ gain(\text{u})}{\sum_{\text{all utterances u}} gain(\text{u})} \quad (18)$$

To conclude, gain measures the strength of the echo, score expresses the individual importance of each unit of analysis and by combining them the proposed method evaluates collaboration with regards to voice intertwining and inter–animation. Figure 33 depicts an example of collaboration assessment for a chat conversation from which intense collaboration zones can be identified, meaningful to the tutor as these areas contain a dense inter–exchange of semantically related utterances between different chat participants. In the upper part of the image there is the conversation graph with the explicit links in red and the implicit ones in green, while in the lower part there is the graphics of the collaboration score for the conversation, both following the conversation timeline (X axis) measured in intervention unique identifiers. It can be seen that our perspective on collaboration correlates with a high distribution of links between utterances of different participants in a short timeframe.

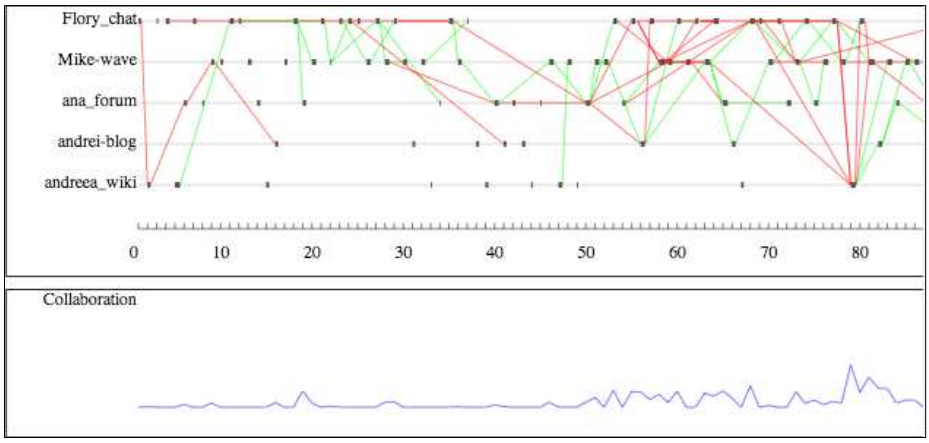


Figure 33. *PolyCAFe* Collaboration evolution within a chat conversation.

All presented computations in terms of intervention scoring and collaboration assessment were later on refined within *ReaderBench* (Dascalu, Trausan-Matu, et al., in press) and are presented in detail in *9 ReaderBench* (3) – Involvement and Collaboration Assessment. In a nutshell, the utterance graph was generalized towards a multi-layered cohesion graph (Trausan-Matu, Dascalu, & Dessus, 2012; Dascalu, Dessus, et al., in press) in which cohesive links are determined through an aggregated similarity measure integrating semantic distances in ontologies (Budanitsky & Hirst, 2006), cosine similarity in latent semantic vector spaces (Landauer & Dumais, 1997) and similarity through topic models from Latent Dirichlet Allocation (Blei et al., 2003). Afterwards, the scoring mechanism was updated in order to best reflect the cohesion-based discourse model (Trausan-Matu, Dascalu, & Dessus, 2012). In the end, the concept of information gain was transposed into the knowledge-building effect, whereas collaboration is regarded as a quantifiable measure of social knowledge-building (Dascalu, Trausan-Matu, et al., in press).

6.4.3 Semantic Extractive Summarization

Based on the utterance graph and on the previously defined analyses mechanisms, each utterance now possesses two scores: a *mark* that reflects its local importance and *gain* (both personal and collaborative) that models the knowledge-building effect. Based on the tight correlation between summary generation and each intervention's importance, we have devised an unsupervised method of extracting chat summaries (Dascalu, Trausan-Matu, & Dessus, 2010a) that takes into consideration the cumulative score and selects the highest ranking interventions, therefore combining collaboration and the marking process of each utterance.

For ease of presentation, we have opted for a predefined percentage of utterances that are displayed in the utterance feedback widget where the user has the possibility to view the entire conversation or only its summary, extremely useful when trying to grasp the key points of the conversation (see Figure 34). Also, in order to add a distinctive sign to the utterances that were selected within the extractive summary, a "star" symbol has been added at the end of each intervention.

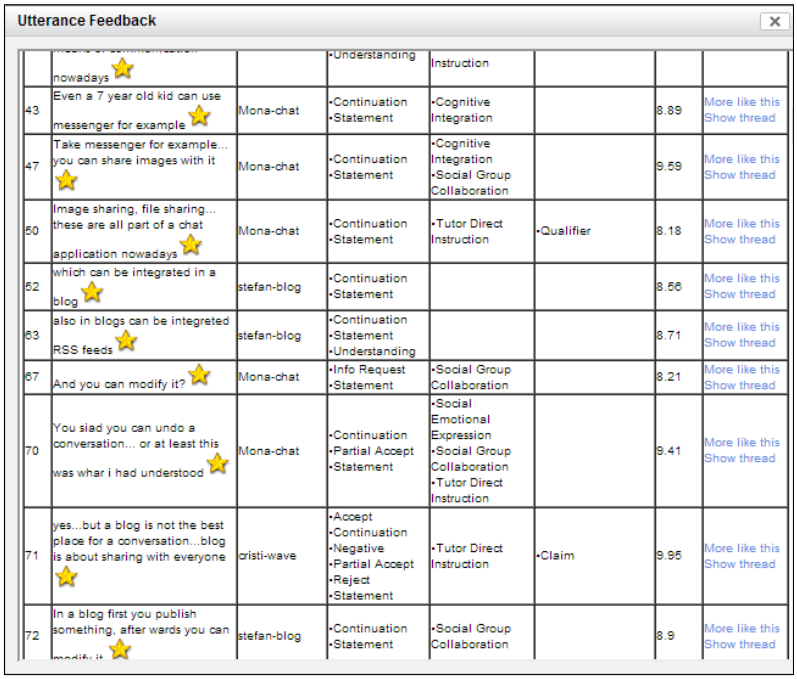


Figure 34. *PolyCAFe* Utterance feedback widget that displays the relevant utterances extracted through the summarization facility.

6.4.4 Semantic Search

A distinctive and yet important component of *PolyCAFe* addresses enhanced search capabilities within a conversation. Starting from a given query, two types of results are being processed: a classification of participants with the criteria of best overall performance with regards to the given query and an utterance list in descending order of relevance scores (see Figure 35).

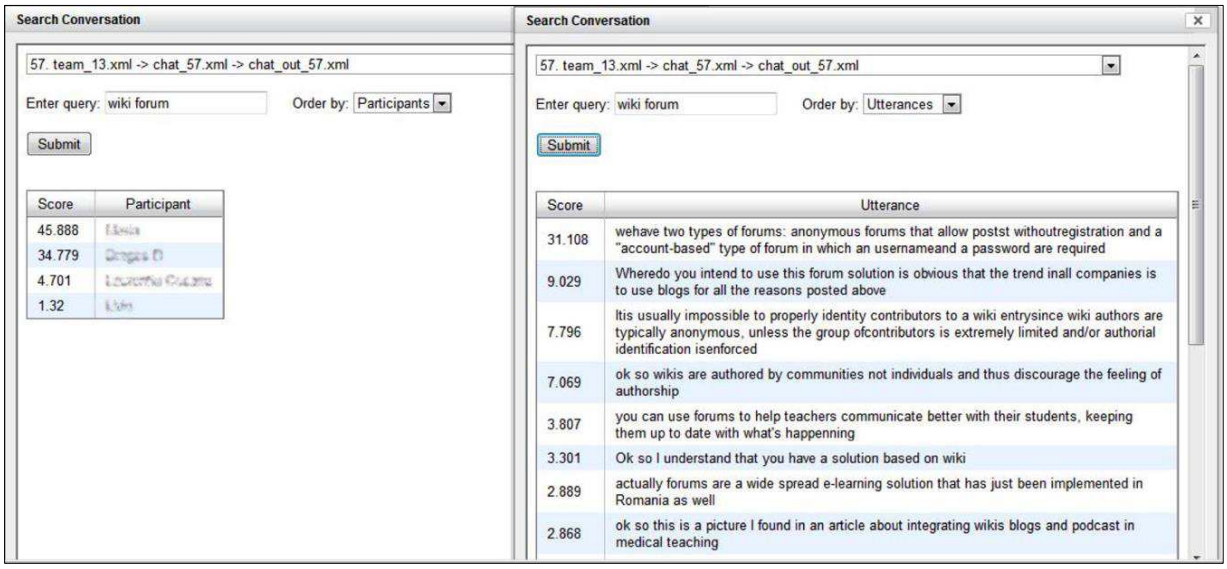


Figure 35. *PolyCAFe* Semantic search – relevance scoring and ordering of: a. Participants; b. Utterances.

In order to evaluate the relevance of each utterance to a given query, three steps are performed. First, the query is lexically and semantically expanded (Voorhees, 1994; Navigli & Velardi, 2003) by using synsets from *WordNet* (Miller, 1995) and by selecting the most promising neighbors from the LSA vector space. For reducing the number of possible newly added items to the query, a threshold and a maximum number of neighbors have been enforced. The final goal is to obtain an expanded list of words and a corresponding query vector that for a small query would have been otherwise biased.

The next step focuses on a morphological analysis, more specifically, on the identification of the occurrences of the expanded query words, in each utterance; in other words, it measures the overlap of concepts between the expanded query and the utterance. In this process, different weights for original inflected forms, concepts added in the process of query expansion and stems are also considered for best reflecting the importance of each utterance.

The third step consists of a semantic assessment by measuring the cosine similarity between the query vector and each utterance's vector; their corresponding vectorial representations are obtained as a normalized sum of concept vectors from the LSA semantic space. The final score is obtained by multiplying the morphological score with the semantic one and with the sum of mark and gain determined in the previous analysis steps. These last two factors were taken into consideration as they incorporate the actual importance and the cumulative knowledge-building effect of each intervention within the discourse. From a cognitive point of view, the displayed final scores should be regarded as a linear distribution of the results' relevance, in terms of its corresponding score (the i^{th} result is X times more relevant than the j^{th} result, where X is $score_i/score_j$). The problem we faced was twofold:

- Simple results ranking would not be sufficient because a high discrepancy between adjacent elements might be encountered (see Figure 35.a in which the concepts in the query are practically addressed only by the first 2 participants).
- Linearly scaling all the scores to a predefined interval (e.g., [0; 1]) would also create confusion because all searches would return at least one entry with the maximum value and no cross-query comparisons would be feasible.

In this context, as the users of the *PolyCAFe* came from a technical background, we opted to actually present them the computed relevance scores (see Figure 35).

6.4.5 Distributed Computing Framework

By taking into consideration the complexity of the computations performed both in *PolyCAFe* and later on in *ReaderBench*, the distributed architecture (Dascalu, Dobre, et al., 2011), although it can be seen as a support function or an external component providing robustness and speedup to the evaluation process, had a major impact while performing analyses on large corpora of documents. The most relevant examples include 1/ building the textual complexity training corpus of more than 1,000 documents from Touchstone Applied Science Associates (TASA) corpus based on the Degree of Reading Power (DRP) score (McNamara et al., in press) and 2/ analyzing a community of practice consisting in more than 400 discussion threads that span across two academic years (Nistor et al., in press).

As an overview with regards to actual data processing, we are, as some authors put it, in an era of "Big Data" (Babu, 2010) where distributed computing has become a commodity. Many enterprises collect large datasets that record customer interactions, product sales, results from advertising campaigns on the Web, etc. Facebook alone collects more than 15 TeraBytes of data each day into its PetaByte-scale data warehouse (Thusoo et al., 2009). Therefore, the ability to perform scalable and timely analytical processing on large datasets to extract useful information is now a critical ingredient for success. Still, cost-effective processing of large datasets is a nontrivial undertaking. Parallel data-flow systems such as *Map-Reduce* (Dean & Ghemawat, 2004; Chu et al., 2007) and *Hadoop* (Borthakur et al., 2011) have recently experienced a surge in popularity (Gates et al., 2009). These systems are increasingly used for data warehousing and analytics, either directly or through the use of a high-level query language that is compiled down to a parallel dataflow graph for execution (Olston, Reed, Srivastava, Kumar, & Tomkins, 2008).

In terms of our conducted analyses, a new approach had to be promoted considering that a multi-threaded, single machine solution is not satisfactory due to the limitations of available resources: at runtime 1 to 5 GB of RAM are allocated (mostly because of POS tagging and statistical parsing) and the process takes 3 to 10 minutes per chat, depending on the actual size of the conversation – from approximately 150 utterances to 450 utterances. Therefore, our aim was to enable the parallel assessment of a corpus on different machines, ensuring also load balancing of task assignments and failover capabilities with the possibility of reassigning failed tasks until the entire corpus is evaluated.

In this context, we too present a solution (Dascalu, Dobre, et al., 2011) that optimizes chat processing using *Map-Reduce*. Our solution, however, is different because we are optimizing an application that is both data and process-intensive. At the same time, the overall task of processing chat conversations suffers from limitations regarding the possible decompositions into concurrent sub-tasks, due to data synchronization and consistency issues. More specifically, the promoted architecture is based on the Replicated Worker paradigm (Garg & Sharapov, 2002) because tasks can be dynamically created as they are generated during the execution of the master/coordinator process. The evaluation processes are identical on each machine and are represented by replicated workers assigned to run on separate physical processors. In addition, the proposed parallel decomposition of the processing operations is also fault-tolerant as the distributed architecture is capable of detecting task execution failures and of re-submitting them to other working nodes. Thus, we optimize the concurrent processing tasks not only by distributing the actual computation on multiple nodes, but also by continuously monitoring and controlling their execution in order to provide a best-effort approach.

Moreover, the proposed architecture is effective because we are dealing with independent tasks, highly computational, with little I/O and little communications between the processes. More generally, we can extrapolate that the presented approach represents a methodology that developers can use to support the optimization through task distribution for other related applications.

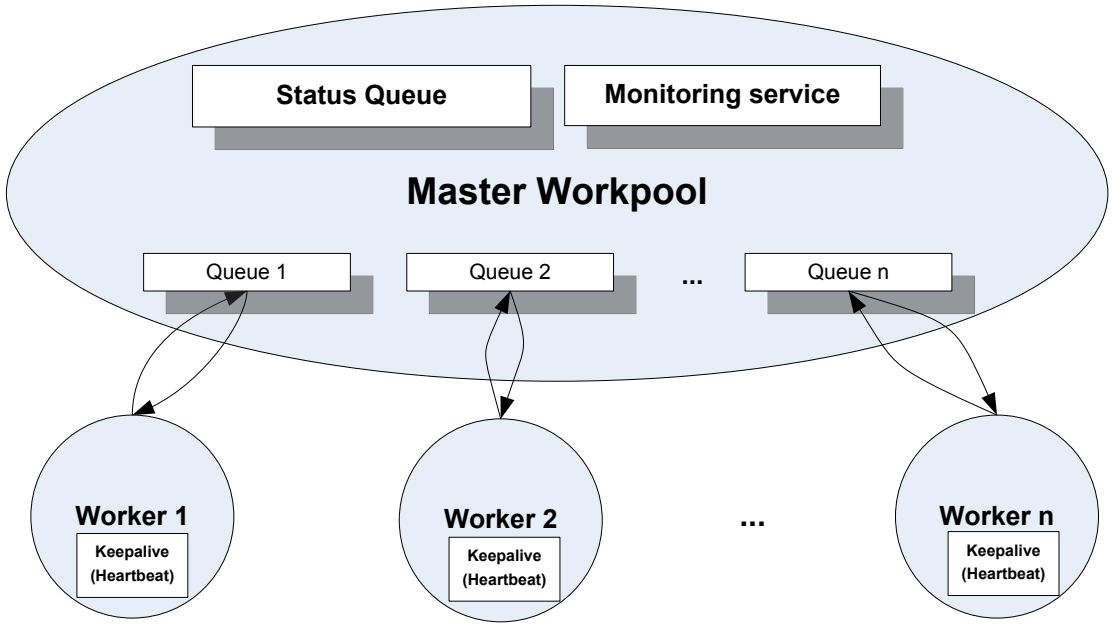


Figure 36. *PolyCAFe* Distributed computing – Replicated workers architecture.

The distributed architecture (see Figure 36) uses a set of distributed work-pools (collection of tasks waiting to be executed by a single worker) that are used for controlling the allocation of tasks to the corresponding workers by using First Come First Served (FCFS) as a planning strategy. When creating a new worker, it registers to the master process and awaits corresponding tasks to be assigned; when it becomes idle, it sends the results to the master and retrieves a new task from its work-pool. After all tasks in the input folder are completed, the master signals all workers for terminations and finishes its own execution. Moreover, the master process also ensures load balancing by assigning new tasks to workers as soon as the current one is finished. Message queues are used for communication, providing an event driven approach and the possibility to send serialized objects between the producer and the consumer of the message. A general message status queue for all workers is used for signaling purposes and task assignments are performed using dedicated queues for each slave.

Failover and redundancy at the worker level are achieved by implementing a modified KAMA – Kaufman Adaptive Moving Average – algorithm (Kaufman, 2005), as the original method considered the noise of the market. The initial approach followed the general trend and continuously adapted the prediction such that, if changes are small and noise is marginal, the predicted values follow the original values very closely. On the other hand, if there are high fluctuations in the received values, KAMA follows with larger distance in order to lessen the number of false predictions. From the market prices to failover and redundancy, the step is performed using a keep-alive thread on each worker that signals its activity periodically and a monitoring service on the master that evaluates whether a worker is still running. If a worker becomes overloaded, the time between two keep-alive messages will increase and the master process must predict correspondingly the next value.

SC (Smoothing Constant) is a standard part of the moving average construction and it determines the level to which the moving average is sensitive to current value swings. SC ranges from 0 to 1: the lower the value, the less sensible the moving average is. In this manner, SC not only follows direction, but also volatility of values. In addition to SC, the following parameters are also considered within our distributed architecture:

- *A* (Actual values) represent the real time of a received keep-alive message;
- *P* (Predicted values) are the master's estimations in terms of the next keep-alive;

- n is the window size or analysis period; in our current implementation the selected value for this parameter is 5;
- α is the Smoothing constant; it is initialized with $\alpha = 2/(n+1)$.

Starting from these definitions, the steps performed in the monitoring process are:

1) *Calculate the ER (Efficiency Ratio)* as direction of the heartbeat or keep-alive messages divided by volatility (see Equations 19, 20 and 21). The more the fluctuation, the greater the rate; on the other hand, if values are constant in time, the direction is 0 and ER becomes also 0.

$$Direction = A_t - A_{t-n} \quad (19)$$

$$Volatility = \sum_{i=0}^{n-1} |A_{t-i} - A_{t-i-1}| \quad (20)$$

$$ER = \frac{|Direction|}{Volatility} \quad (21)$$

2) Using the shortest and the longest moving average, we next determine the SC of these averages (see Equations 22 and 23). The used window sizes range from 2 messages to 10 messages.

$$SC = ER \times (Fasc_{SC} - Slow_{SC}) + Slow_{SC} \quad (22)$$

where

$$Fast_{SC} = \frac{2}{2+1}, Slow_{SC} = \frac{2}{10+1} \quad (23)$$

KAMA shortens or extends the time period used for computing the moving average according to the conditions that prevail within the architecture. KAMA becomes more sensible or robust tightly connected with the frequency of the heartbeat messages received.

3) The final smoothing constant (see Equation 24) is obtained by squaring SC in order to make it less sensible and to avoid a zigzag evolution:

$$\alpha = SC^2 \quad (24)$$

4) The final KAMA computation looks similar to the Exponential Moving Average (EMA) calculation by approximating the new predicted value for receiving a new heartbeat message based on the last received heartbeat value and the previous prediction (see Equation 25).

$$P_t = \alpha \times A_{t-1} + (1 - \alpha) \times P_{t-1} \quad (25)$$

KAMA belongs to less known moving averages and its main advantage is that it takes into consideration not just the direction, but also the heartbeat time volatility. KAMA adjusts its length according to the prevailing keep-alive signaling conditions and informs us about the trends during parallel conversations evaluation.

In the end, a *suspicion level* (see Equation 26) ranging in [0; 1] is determined for each worker. While a small value, close to 0, represents a fully functional and working process, the closer the value gets to 1, the greater the probability of a malfunction. Therefore, with each missed or delayed heartbeat message, the level of doubt increases.

$$sl(t) = \frac{e^{\beta(t-1)} - 1}{e^{\beta(t-1)} + 1}, t = \frac{t_{now}}{t_{predicted}} \quad (26)$$

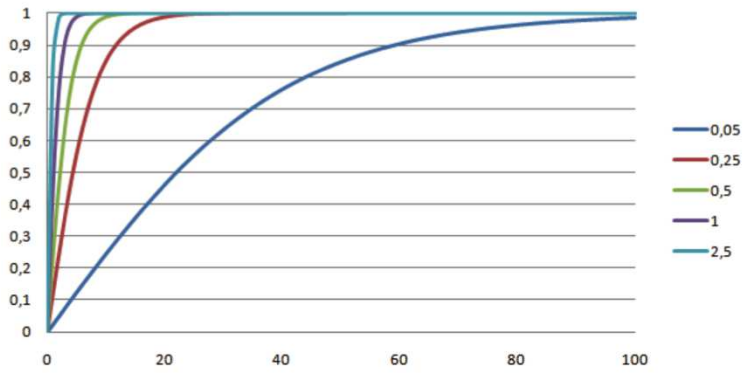


Figure 37. *PolyCAFe* Distributed computing – Suspicion level evolution for different β values.

Starting from the evolution of the suspicion levels depicted in Figure 37, the imposed threshold was *0.8*. After surpassing this value, the process is considered malfunctioning/unavailable and the task is reassigned, later to be solved by a functional worker. Fine-tuning of the function can be obtained by adjusting the values of β. Therefore, for smaller values, more time representing a greater number of missed keep-alive messages is allowed to pass until the master process declares the process dead. The value for β used in current tests was of 0.5, enabling a maximum of 5 heartbeat messages to be missed. Because the only single point of failure in the described architecture is the master process,

checkpoints are made after each completed task, enabling easy restore of the corpus assessment process in the eventuality the coordinator fails during its execution. Fast resume is possible in the case of an execution error in the master process by saving a snapshot each time a task is successfully completed.

In terms of performance, the speedup of the architecture is considerable and can be estimated to the actual number of workers multiplied by a factor of .94 (Dascalu, Dobre, et al., 2011); this value expresses latency with regards to communications and internal context switching due of other running threads. Therefore, we can conclude that the main benefits of our approach are robustness, adaptability to worker heartbeat messages and scalability allowing virtually an unlimited number of workers to be assigned on individual tasks.

6.5 Validation of *PolyCAFe*

The validation of *PolyCAFe* consisted of two rounds presented in detail in this section. Whereas the first run showed promising results, the sample of participants was quite limited; therefore, a second round of validation experiments was conducted, followed by a thorough verification of *PolyCAFe*'s results.

6.5.1 First Validation

A first validation (Rebedea et al., 2010; Dascalu, Rebedea, et al., 2011) has been performed during the Human Computer Interaction course, in the academic year 2009-2010, at the Computer Science department involving 9 senior (4th year) students and 5 tutors that used *PolyCAFe* for analyzing the conversations and providing feedback to the students. The experiment was structured in the following manner: students had to read online and printed materials on a given topic (CSCL technologies) and then they had a debate using *ConcertChat* (Holmer et al., 2006) in two small groups of 4-5 students. After the debate, they used *PolyCAFe's* feedback to understand their involvement in the conversation and what could have been improved. Two tutors monitored this activity, provided help to the students and took notes regarding the asked questions, comments and the actual behavior when interacting with the widgets. As the students were encouraged to think aloud when using the system, these data were useful for identifying the main problems of the software. Besides thinking aloud, the students were also asked to use a document where they registered what they considered misleading in the feedback returned by the system. This activity

lasted 90-120 minutes and was followed by a questionnaire with 32 validation statements with answers on a 5-level Likert scale (1-strongly disagree – 5-strongly agree), grouped in five categories: Pedagogic effectiveness, Efficiency, Cognitive load, Usability and Satisfaction. Afterwards, a focus group with all students was conducted in order to find the most important advantages and disadvantages, plus suggestions for improvements.

On the other hand, tutors were asked to provide feedback to a conversation using *PolyCAFe* and to another chat without the system. After this step, they were invited to answer a questionnaire with 35 validation statements using the same scale and categories as the one for the students. Then all tutors took part in a focus group where they were invited to share their points of view about each feature's utility, about the reliability of the feedback and what improvements they envisioned.

Overall, all students and tutors considered the feedback provided by *PolyCAFe* to be useful and relevant for their task (Rebedea et al., 2010), but the opinion of students was divided as just 63% of them considered that the feedback was helpful in improving their learning experience. One explanation for this result might be the fact that the students have never used *PolyCAFe* prior to the validation experiment and it might have been difficult for them to understand how to use all the provided facilities. On the other hand, the validation experiments highlighted that the usability of the widgets could also be improved and thus, their relevance and user acceptance might increase.

Table 16 (Dascalu, Rebedea, et al., 2011) presents the aggregated validation results on all the five categories for both tutors and students. It is clear that all the tutors found *PolyCAFe* efficient for their task, as it helps them reduce the time needed for writing feedback for students and it improves the quantity and consistency of this feedback among tutors. Moreover, it is easily noticeable that the student results are worse for all categories than the ones of the tutors. The lowest score was obtained for cognitive load, showing that the users had some problems accommodating to *PolyCAFe* on their first use. In addition, the results show that more than a quarter of the learners are not satisfied by the system and the main presented reason was that the students didn't trust the statistical results displayed as they considered some indicators to be misleading.

Table 16. *PolyCAFe* First validation results per category using the 5-level Likert scale (1-strongly disagree – 5-strongly agree).

	Tutor		Student	
Validation statement category	Average score	Agreement	Average score	Agreement
Pedagogic effectiveness	4.11	83%	3.94	77%
Efficiency	5.00	100%	4.22	78%
Cognitive load	4.60	100%	3.56	56%
Usability	4.36	93%	4.11	81%
Satisfaction	4.57	91%	3.89	72%
Total	*4.53*	*93%*	*3.94*	*73%*

While taking a closer look at the questionnaires, the tutors had agreed with all but one statement with average scores between 4.11 and 5.00, while the students had agreed with 27 out of the 32 statements, with average scores between 3.56 and 4.22. As it can be noted, there are considerable differences between the students' results and those of the tutors. However, possible explanations for these discrepancies might be: 1/ the tutors were more familiarized with the system as they had used it prior to analyze other conversations or 2/ the tutors considered the system helpful as *PolyCAFe* provides them feedback more quickly and as it improved the time required for the analysis by up to 50%, and in a reliable manner.

All the identified aspects within this preliminary validation experiment were used to increase the reliability and the usability of *PolyCAFe* and were treated in detail in a second, more elaborated, validation experiment.

6.5.2 Second Validation

After the first validation pilot showed that the system was efficient and effective for both learners and tutors, a new validation (Rebedea, Dascalu, Trausan-Matu, Armitt, et al., 2011) was undertaken to further study the effects of using an improved version of *PolyCAFe*, with a larger group of students. The new experiment was integrated as a learning task and assignment for a group of senior year undergraduate students studying Human-Computer Interaction, academic year 2010-2011. A total of 35 students have been engaged in the experiment for several weeks: 25 students were part of the experimental group and 10 students were assigned to the control group. The only difference between the experimental and control group is that the latter did not receive any feedback from *PolyCAFe*,

but only from the tutors. The learners were divided into groups of 5 students, thus having 5 experimental and 2 control groups, and were given two successive chat assignments related to CSCL technologies (chat, blog, wiki, forums and *Google Wave*) to debate using *ConcertChat* (Holmer et al., 2006).

In the first assignment, the experimental group was asked to use *PolyCAFe* to get feedback, while the control group did not use the system. The use of *PolyCAFe* for the second assignment was not mandatory, so the learners had an option to use the system only if they considered it would be useful for them. The tutors had to provide manual feedback to each of the students involved in the chat conversations for the first assignment. Each tutor assessed at least one conversation without using *PolyCAFe* and one conversation using the system. With regard to the second assignment, no manual feedback was provided, only the outputs of *PolyCAFe*.

At the end of the validation experiment, all the students and tutors had to answer a questionnaire and participated in focus groups and interviews. The results of the experiment have been devised into several validation topics: tutor efficiency, quality and consistency of the automatic feedback, making the educational process transparent, quality of educational output, motivation for learning, etc. All these topics have been validated, but only three were analyzed in extenso as they were considered central to the educational scenario and to the usage of the system for similar CSCL tasks:

- *VT1*: Tutors/facilitators spend less time preparing feedback for learners compared with traditional means (tutor efficiency);
- *VT2*: Learners perceive that the feedback received from the system contributes to informing their study activities (quality and consistency of automatic feedback);
- *VT3*: Learner performance in online discussions is improved in the areas of content coverage and collaboration, when using *PolyCAFe* (quality of educational output).

In order to validate *tutor efficiency*, several methods have been used: measurements, questionnaires and the answers to the interviews. Overall, all show a good consensus of the 6 tutors with regard to the efficiency of using *PolyCAFe*, with averages over 4.5 and agreement factors of over 83% for all the validation statements. In addition to the first validation experiments, time measurements for preparing the feedback by tutors were also used. Thus, 4 tutors analyzed each chat conversation, 2 using *PolyCAFe* and the other without using the system. This data has been compared for all the 7 chats resulted for the first assignment. The average time needed to prepare feedback without

PolyCAFe was of 84 minutes, with a standard deviation of 15 minutes, while the average time required for providing feedback with *PolyCAFe* was of 55 minutes with a standard deviation of 20 minutes. These results show a significant average time reduction for a single chat conversation: (84 – 55)/84 = 35%. However, as the standard deviation has increased, it also demonstrates that not all tutors managed to use the software efficiently.

Quality and consistency of the automatic feedback has been validated using questionnaires for the experimental group of 25 students, plus system logging. The statements focused on the accuracy, the relevance, the usefulness and consistency of the provided feedback – all were validated with agreement factors between 60-80% and means between 3.70 and 4.00. The system logging utilities monitoring student access to *PolyCAFe* have shown that for the whole period of the validation experiment there have been 285 visits and 1,447 page-views, resulting in more than 40 page views in average per student. Therefore, the students have been actively using the system in order to reflect on their activity in specific chat conversations.

The last topic, *the quality of the educational output*, was validated through measurements computed by *PolyCAFe* for the second chat assignment, as a comparison between the experimental groups versus the control groups: the most important concepts in the conversation and their score, the average grade for utterances throughout the entire conversation, the number of interventions and the density of implicit and explicit links between utterances. However, only the average scores of utterances and the quantitative estimation of collaboration through the density of links (average number of links/utterance) showed a noticeable increase between the two groups: 6.8% for the number of utterances and 29% for the estimation of collaboration, in favor of the experimental group.

6.5.3 Participant Ranking Verification

For all the chats of the first assignment, the tutors had to rank the participants according to the importance they had throughout the conversation, taking into account the content of their interventions, their involvement and the degree of collaboration with the other participants. Therefore, each tutor assigned a rank from 1 to 5 to each participant; additionally, each participant was ranked by all other participants pertaining to the same conversation These results were then compared with the automatic ranking of participants provided by *PolyCAFe* (see Table 17) (Trausan-Matu et al., 2011).

Table 17. *PolyCAFe* Sample of participant rankings for a single chat conversation.

Rank	Stud. 1	Stud. 2	Stud. 3	Stud. 4	Stud. 5	Stud. avg.	Tutor 1	Tutor 2	Tutor avg.	*PolyCAFe*
Student 1	-	2	2	1	1	2	4	4	4	4
Student 2	2	-	3	2	2	3	1	2	1-2	2
Student 3	3	3	-	3	4	4	5	5	5	5
Student 4	1	1	1	-	3	1	2	1	1-2	1
Student 5	4	4	4	4	-	5	3	3	3	3

By analyzing all 7 conversations, *PolyCAFe* achieved an excellent precision and correlation with the average tutor scores (r = .94 and P = 77%) and good results with the average student scores (r = .84 and P = 66%) (see Table 18) (Rebedea, Dascalu, Trausan-Matu, Armitt, et al., 2011).

Table 18. *PolyCAFe* Comparison of average participant rankings.

Comparison	Correlation	Precision	Average error
Tutors – System	.94	77%	0.23
Students – System	.84	66%	0.43
Tutors – Students	.84	71%	0.40

Moreover, although the average rankings of the students, the tutors and *PolyCAFe* are quite well correlated one with the other, individual basis correlations drop dramatically due to the fact that a simple inversion in a series of 5 elements (the number of participants per conversion) changes the trend and therefore drastically diminishes inter-rater correlations. Nevertheless, the results encourage us to conclude that, although the grading or ranking criteria of tutors and students are not the same, the system is well correlated with the average values, thus being more objective.

6.6 Conclusions and Transferability

In contrast to the previously implemented systems, *PolyCAFe* was particularly designed for providing comprehensive feedback to both tutors and learners using chat conversations for collaborative assignments. From the learner perspective, the displayed indicators and the provided textual feedback were the main benefits for improving the student's learning activities through collaboration. As for the tutors, they considered that reducing the time needed to provide manual feedback to their students was the greatest advantage of using *PolyCAFe*.

Overall, from a technical viewpoint, *PolyCAFe* can be considered a complex learning analytics tool for online conversations that underpins the dialogic and polyphonic theories and that employs NLP and SNA techniques in order to discover implicit relations between utterances. Afterwards, an utterance graph is built, used later on for evaluating utterances, participants' involvement and collaboration. Moreover, the validation experiments performed in a formal, academic environment have shown the acceptance and the usefulness of *PolyCAFe* by both learners and tutors.

With regards to transferability in different educational scenarios, the following dimensions must be taken into consideration: *domain*, *language* and the *learning task*. As the system was developed for English only, in order to ensure *language transferability* new linguistic tools must be integrated for each new language (e.g., the entire NLP processing pipe, lexicalized ontology, adjacency pairs and other linguistic patterns).

Domain transferability is mostly concerned with the existence of a large corpus of text documents, relevant to the task at hand, that are required in order to build the LSA vector space. In addition, all domains, where textual descriptions of descriptive knowledge are used, are well suited (e.g., *PolyCAFe* was successfully used on Medicine discussion forums). On the contrary, there are domains where PolyCAFe is not well suited due to the need of graphical elements or images or general discussions, without a clear focus and for which is difficult to build a relevant LSA space.

Moreover, from a *pedagogical point of view*, *PolyCAFe* can be used in a wide variety of collaborative contexts: role-based discussions and debates, open argumentations, problem solving (in mathematics and design or any domain specific task) or creative discussions (brainstorming) (Trausan-Matu, 2010b). More specifically, we envision the following contexts: revising exams and discussions on given topics, finding collaborative solutions to problems that can be described without the importance of a sequence of steps (PBL) or further investigation of a given topic of interest to the learner (Self-Regulated Learning). On the other hand, *PolyCAFe* is not suitable for learning scenarios in which collaboration is not required, nor encouraged, or for settings that involve scripted collaboration following a set of exhaustive instructions.

7 *ReaderBench* (1) – Cohesion-based Discourse Analysis and Dialogism

7.1 Overview of *ReaderBench*

In contrast to the previous systems that focused on analyzing CSCL conversations, *ReaderBench* (Dascalu, Dessus, et al., in press; Dascalu, Trausan-Matu, et al., in press) (see Table 19) addresses a wider spread of activities and can be used within more complex educational scenarios (see *10.3* Educational Implications). Nevertheless, *A.S.A.P.*, *Ch.A.M.P* and *PolyCAFe* have provided valuable insight and some features were reused, of course with the necessary improvements. Moreover, an important aspect when considering *ReaderBench* is the French National Agency of Research project DEVCOMP ANR-10-BLAN-1907 in which the gold standard consisting of learner materials and the assessment of reading strategies expressed in pupils' verbalizations have been developed.

Table 19. *ReaderBench* (1) Traceability matrix of provided functionalities and integrated tools.

	Tools								
Functionality	Patterns	IR	NLP pipe	*WordNet/ WOLF*	LSA	LDA	SVM	SNA	Distributed computing
Cohesion graph and document/conversation visualization	✓		✓	✓	✓	✓		✓	
Topics modeling		✓		✓	✓	✓			
Dialogical perspective	✓	✓	✓	✓	✓	✓			
Reading strategies identification	✓	✓	✓	✓	✓	✓			
Textual complexity assessment		✓	✓	✓	✓	✓	✓		✓
Participant involvement evaluation			✓	✓	✓	✓		✓	✓

Intervention scoring	✓	✓	✓	✓	✓	✓	✓
Collaboration assessment	✓		✓	✓	✓		✓
Semantic extractive summarization	✓		✓	✓	✓	✓	

In addition, *ReaderBench* can be considered a novel approach as it introduced a generalized model for assessment based on the cohesion graph (Trausan-Matu, Dascalu, & Dessus, 2012) (see *7.2* Cohesion-based Discourse Analysis), applicable to both plain essay- or story-like texts (Dascalu, Dessus, et al., in press), and CSCL conversations (Dascalu, Trausan-Matu, et al., in press), in particular chats or forum discussion threads (Nistor, Baltes, et al., submitted; Nistor, Dascalu, et al., submitted). Figure 38 and Table 20 present the information flow and actual computational steps, main evaluation steps and the mapping to the corresponding section containing a detailed description of the undergone processes.

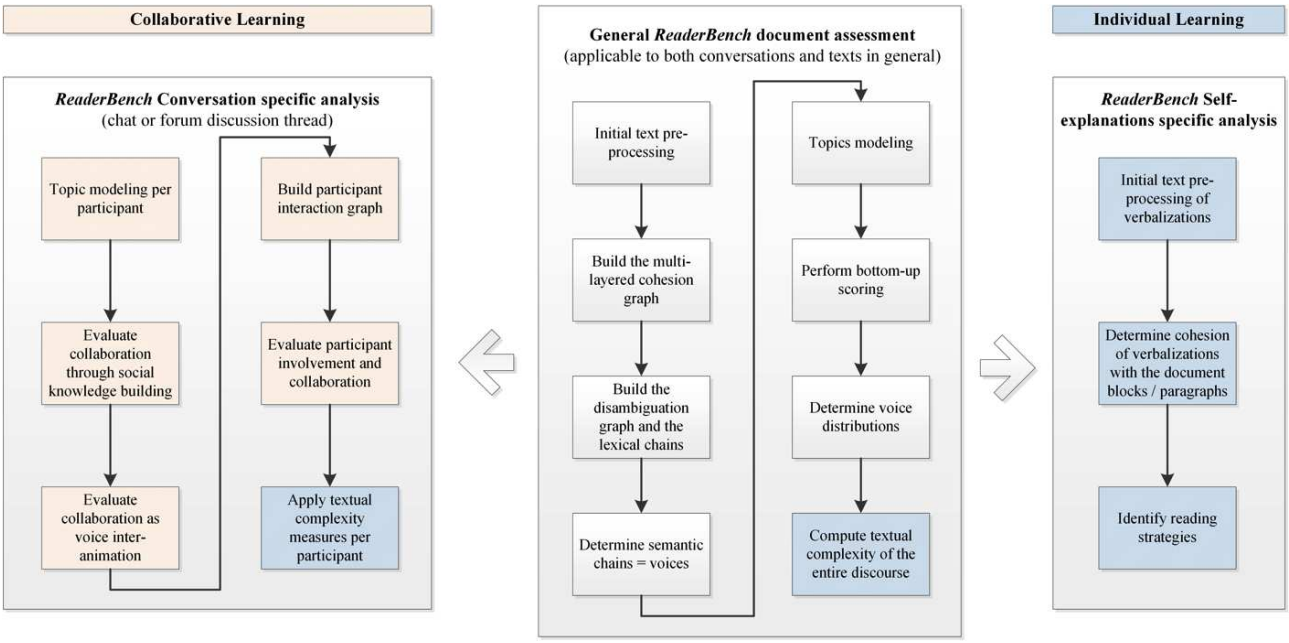


Figure 38. *ReaderBench* (1) General workflow.

Table 20. *ReaderBench* (1) Mapping between workflow steps and corresponding detailed descriptions.

No.	Workflow step	Section with detailed description
General analysis		
1	Initial text pre-processing	*7.2* Cohesion-based Discourse Analysis
2	Building the multi-layered cohesion graph	*7.2* Cohesion-based Discourse Analysis
3	Building the disambiguation graph and the lexical chains	*4.3.1* Semantic Distances and Lexical Chains
4	Determine semantic chains = voices	*7.5* Dialogism and Voice Inter-Animation
5	Topics modeling	*7.3* Topics Extraction
6	Perform bottom-up scoring	*7.4* Cohesion-based Scoring Mechanism
7	Determine voice distributions	*7.5* Dialogism and Voice Inter-Animation
8	Compute complexity of the entire discourse	*8.2* Textual Complexity Analysis Model
Self-explanations specific analysis		
9*	Initial text pre-processing of verbalization	*7.2* Cohesion-based Discourse Analysis
10*	Determine cohesion of verbalizations with the document blocks/paragraphs	*8.1.3* Reading Strategies Identification Heuristics
11*	Identify reading strategies	*8.1.3* Reading Strategies Identification Heuristics
Conversation (chat or forum discussion thread) specific analysis		
9*	Topic modeling per participant	*9.1* Participant Involvement Evaluation
10*	Evaluate collaboration through social knowledge-building	*9.2.1* Social Knowledge-Building Model
11*	Evaluate collaboration as voice inter-animation	*9.2.2* Dialogical Voice Inter-Animation Model
12*	Build participant interaction graph	*9.1* Participant Involvement Evaluation
13*	Evaluate participant involvement and collaboration	*9.1* Participant Involvement Evaluation
14*	Apply textual complexity measures per participant	*8.2* Textual Complexity Analysis Model

As an overview, *ReaderBench* consists of a document core assessment engine presented in this chapter, with extensions addressing the particularities of the two learning scenarios presented in *10.3.1* Envisioned Educational Scenarios focused on *individual* (see *8 ReaderBench* (2) – Individual Assessment through Reading Strategies and Textual Complexity) and *collaborative* (see *9 ReaderBench* (3) – Involvement and Collaboration Assessment through Cohesion and Dialogism)

learning. Therefore, starting from a cohesion-based representation of discourse, ReaderBench addresses three major issues: 1/ the identification of reading strategies, 2/ textual complexity assessment and 3/ involvement and collaboration in CSCL conversations. In addition to all previously developed systems, *ReaderBench* has increased multi-lingual support and integrates specific NLP tools for both French and English.Due to its complexity, we opted for splitting the presentation of *ReaderBench* into three adjacent chapters, each focused on a specific purpose, but all providing in the end interdependent functionalities. Additional workflows and print-screens are presented in Appendix A – Document Workflow and Additional Print-screens and Appendix B – Verbalization Workflow and Additional Print-screens.

7.2 Cohesion-based Discourse Analysis

Text cohesion, viewed as lexical, grammatical and semantic relationships that link together textual units (see *2.1.1* Coherence and Cohesion and *4.1* Measures of Cohesion and Local Coherence), is defined within our implemented model in terms of: 1/ the *inverse normalized distance between textual elements*; 2/ *lexical proximity* that is easily identifiable through identical lemmas and semantic distances (Z. Wu & Palmer, 1994; Leacock & Chodorow, 1998) within ontologies; 3/ *semantic similarity* measured through Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) (see *4.3* Natural Language Processing Techniques). Additionally, specific natural language processing techniques (Manning & Schütze, 1999) are applied to reduce noise and improve the system's accuracy: spell-checking (optional) (Alias-i, 2008; McCandless, Hatcher, & Gospodnetic, 2010), tokenizing, splitting, part of speech tagging (Toutanova & Manning, 2000; Toutanova, Klein, Manning, & Singer, 2003), parsing (Klein & Manning, 2003; Green, de Marneffe, Bauer, & Manning, 2010), stop words elimination, dictionary-only words selection, stemming (Porter & Boulton, 2002), lemmatizing (Jadelot, Mangeot, Petitjean, & Salmon-Alt, 2006), named entity recognition (Finkel, Grenager, & Manning, 2005) and co-reference resolution (Raghunathan et al., 2010; Lee et al., 2011; Lee et al., 2013) (see Figure 39).

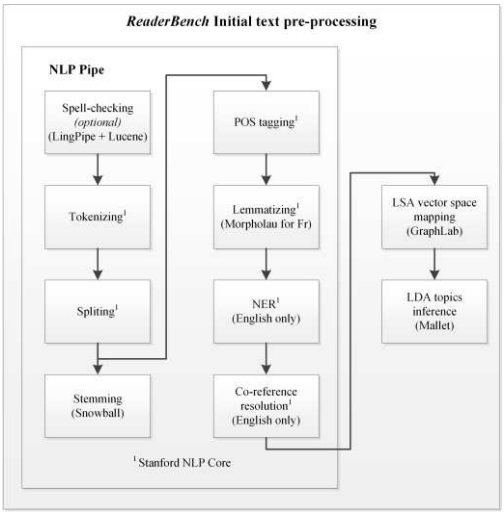


Figure 39. *ReaderBench* (1) Initial text pre-processing (the extended NLP pipe).

In order to provide a multi-lingual analysis platform with support for both English and French, *ReaderBench* integrates both *WordNet* (Miller, 1995) and a transposed and serialized version of *Wordnet Libre du Français* (WOLF) (Sagot, 2008; Sagot & Darja, 2008). Due to the intrinsic limitations of WOLF, in which concepts are translated from English while their corresponding glosses are only partially translated, making a mixture of French and English definitions, only three frequently used semantic distances were applicable to both ontologies: path length, Wu-Palmer (Z. Wu & Palmer, 1994) and Leacock-Chodorow's normalized path length (Leacock & Chodorow, 1998).

Afterwards, LSA and LDA semantic models (see *4.3.2* Semantic Similarity through Tagged LSA and *4.3.3* Topic Relatedness through Latent Dirichlet Allocation) were trained using three specific corpora: "*TextEnfants*" (Denhière, Lemaire, Bellissens, & Jhean-Larose, 2007) (approx. 4.2M words), "*Le Monde*" (French newspaper, approx. 24M words) for French, and Touchstone Applied Science Associates (TASA) corpus (approx. 13M words) for English. Moreover, improvements have been enforced on the initial models: the reduction of inflected forms to their lemmas, the annotation of each word with its corresponding part of speech through a NLP processing pipe (only for English as for French it was unfeasible to apply to the entire training corpus due to the limitations of the Stanford Core NLP in parsing French) (Manning & Schütze, 1999; Wiemer-Hastings & Zipitria, 2000; Lemaire, 2009; Dascalu, Trausan-Matu, et al., 2010b), the normalization of occurrences through the use of term frequency-inverse document frequency (*Tf-Idf*) (Manning & Schütze, 1999) and distributed computing, enabling a concurrent and parallel execution of tasks, for increasing speedup (McCallum, 2002; Low et al., 2010; Low et al., 2012).

LSA and LDA models extract semantic closeness relations from underlying word co-occurrences and are based on the bag-of-words hypothesis (see *4.3.2* Semantic Similarity through Tagged LSA and *4.3.3* Topic Relatedness through Latent Dirichlet Allocation). Our experiments have proven that LSA and LDA models can be used to complement one other, in the sense that underlying semantic relationships are more likely to be identified, if both approaches are combined after normalization. Therefore, LSA semantic spaces are generated after projecting the arrays obtained from the reduced-rank Singular Value Decomposition of the initial term-doc array and can be used to determine the proximity of words through cosine similarity (Landauer & Dumais, 1997; Landauer, Foltz, et al., 1998). From a different viewpoint, LDA topic models provide an inference mechanism of underlying topic structures through a generative probabilistic process (Blei et al., 2003). In this context, similarity between concepts can be seen as the opposite of the Jensen-Shannon dissimilarity (Manning & Schütze, 1999) between their corresponding posterior topic distributions.

From a computational perspective, the LSA semantic spaces were trained using a Tagged LSA engine (Dascalu, Trausan-Matu, et al., 2010b) that preprocesses all training corpora (stop-words elimination, POS tagging, lemmatization) (Wiemer-Hastings & Zipitria, 2000; Lemaire, 2009), applies *Tf-Idf* and uses a distributed architecture (Low et al., 2010; Low et al., 2012) to perform the Singular Values Decomposition. With regards to LDA, the parallel topics model used iterative Gibbs sampling over the training corpora (McCallum, 2002) with 10,000 iterations and 100 topics, as recommended by Blei et al. (2003). Overall, in order to better grasp cohesion between textual fragments, we have combined information retrieval specific techniques, mostly reflected in word repetitions and normalized number of occurrences, with semantic distances extracted from ontologies or from LSA- or LDA-based semantic models.

In order to have a better representation of discourse in terms of underlying cohesive links, we introduced a cohesion graph (Trausan-Matu, Dascalu, & Dessus, 2012; Dascalu, Trausan-Matu, et al., in press) that can be seen as a generalization of the previously proposed utterance graph (Trausan-Matu, Stahl, et al., 2007; Rebedea & Trausan-Matu, 2009; Trausan-Matu & Rebedea, 2010; Rebedea, Dascalu, Trausan-Matu, & Chiru, 2011). More formally, we are building a multi-layered mixed graph consisting of three types of nodes (see Figure 40 and Figure 41):

- A central node, the *document* that represents the entire reading material or the conversation.

- *Blocks*, a generic entity that can reflect paragraphs from the initial text or utterances/interventions in chat conversations or forum threads.
- *Sentences*, the main units of analysis, seen as collections of words and grammatical structures obtained after the initial NLP processing.

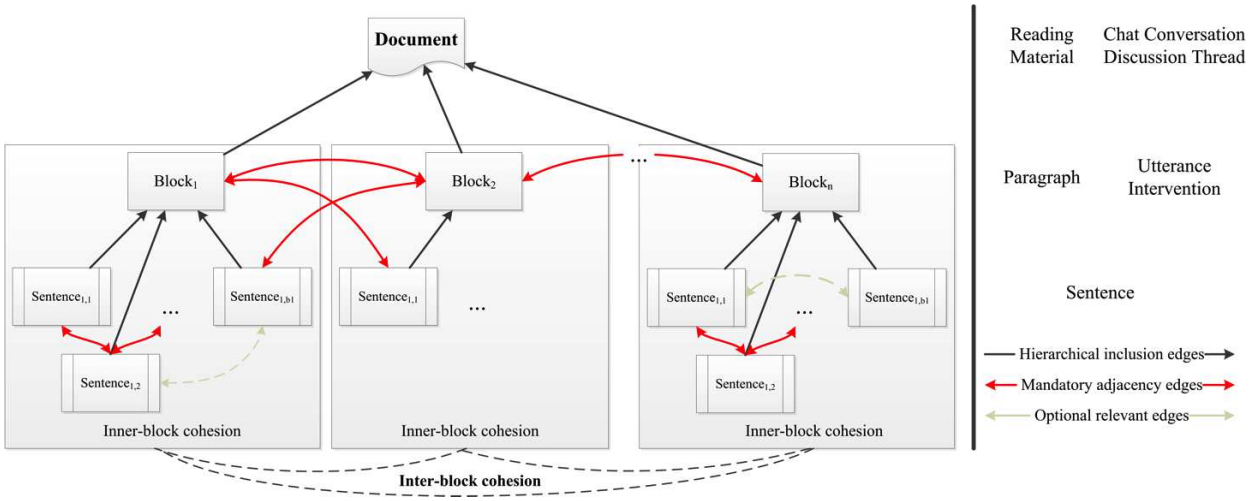


Figure 40. *ReaderBench* (1) Cohesion Graph.

In terms of *edges*, *hierarchical links* are enforced through inclusion functions (sentences within a block, blocks within the document) and two types of links are introduced between analysis elements of the same level: *mandatory* and *relevant links*.

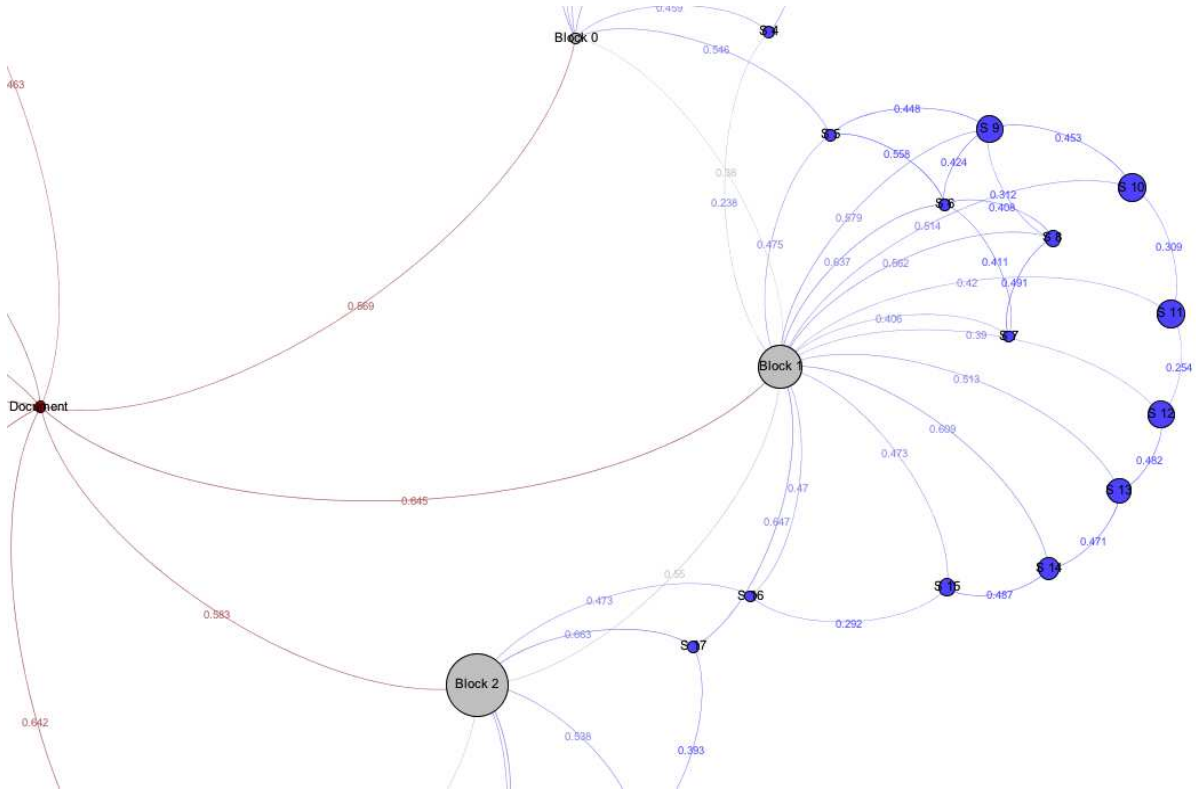


Figure 41. *ReaderBench* (1) Generated partial view of a Cohesion Graph.

Mandatory links are established between adjacent blocks or sentences and are used for best modeling the information flow throughout the discourse, therefore making possible the identification of cohesion gaps. In the case of chat conversations, if the user adds explicit links within the user interface of an ongoing discussion – in our case, *ConcertChat* (Holmer et al., 2006) –, these links are also added to the cohesion graph and considered mandatory. In the case of forum thread discussions, a similar link is added between the reply post and the corresponding initial intervention.

Moreover, adjacency links are enforced between the previous block and the first sentence of the next block and, symmetrically, between the last sentence of the current block with the next block. This is performed in order to ensure cohesiveness between structures at different levels within the cohesion graph, disjoint with regards to the previous inclusion function, and for augmenting the importance of the first/last sentence of the current block, in accordance with the assumption that topic sentences are usually at the beginning of the paragraph and in most cases ensure a transition from the previous paragraph (Abrams, 2000).

Additional optional *relevant links* are added to the cohesion graph for highlighting fine-grained and subtle relations between distant analysis elements. In our experiments, the use as threshold of the sum of mean and standard deviation of all cohesion values from within a higher-level analysis element provided significant additional links into the proposed discourse structure. Also, for chat conversations, the search space for significant implicit cohesive links has been limited to 20 adjacent utterances, as previous experiments demonstrated that the probability to have explicit links outside this window span is close to 0 (Rebedea, 2012).

In contrast, as cohesion can be regarded as the sum of semantic links that hold a text together and give it meaning, the mere use of semantically related words in a text does not directly correlate with its complexity. In other words, whereas cohesion in itself is not enough to distinguish texts in terms of complexity, the lack of cohesion may increase textual complexity, as a text's proper understanding and representation become more difficult to achieve. In order to better highlight this perspective and emphasize the theoretical aspects presented in *2.1.3* Cohesion and Coherence versus Textual Complexity, two measures for textual complexity were defined, later to be assessed: *inner-block cohesion* as the mean value of all the links from within a block (adjacent and relevant links between sentences) and *inter-block cohesion* that highlights semantic relationships at global document level.

Our aim was to provide a generalized and customizable model for assessing different types of discourses in terms of cohesion: multi-participant chat conversations or forum discussion threads, on one hand, and reading materials, meta-cognitions or general texts, on the other. More specific to CSCL conversations, we opted for Dong's (2005) perspective of separating utterances based on turn-taking events between speakers. Although most participants' interventions consist of solely one sentence and elliptical expressions are quite frequent, we preferred to create an extensible model, easily adjustable to different types of inputs.

As *validation*, we have used 10 stories in French for which sophomore students in educational sciences (French native speakers) were asked to evaluate the semantic relatedness between adjacent paragraphs on a Likert scale of [1; 5]; each pair of paragraphs was assessed by more than 10 human evaluators for limiting inter-rater disagreement. Due to the subjectivity of the task and the different personal scales of perceived cohesion, the average values of intra-class correlations per story were *ICC-average measures* = .493 and *ICC-single measures* = .167. In the end, 540 individual cohesion scores were aggregated and then used to determine the correlation between different semantic measures and the gold standard. On the two training corpora used (*Le Monde* and *TextEnfants*), the correlations were: *Combined-Le Monde* (r = .54), *LDA-Le Monde* (r = .42), *LSA-Le Monde* (r = .28), *LSA-TextEnfants* (r = .19), *Combined-TextEnfants* (r = .06), *Wu-Palmer* (r = -.06), *Path Similarity* (r = -.13), *LDA-TextEnfants* (r = -.13) and *Leacock-Chodorow* (r = -.40). All these correlations are non-significant, but the inter-rater correlations are on a similar range and are smaller than the *Combined-Le Monde* score.

The previous results show that the proposed combined method of integrating multiple semantic similarity measures outperforms all individual metrics, that a larger corpus leads to better results and that Wu-Palmer, besides its corresponding scaling to the [0; 1] interval (relevant when integrating measurements with LSA and LDA), behaves best in contrast to the other ontology based semantic distances. Moreover, the significant increase in correlation between the aggregated measure of LSA, LDA and Wu-Palmer, in comparison to the individual scores, proves the benefits of combining multiple complementary approaches in terms of the reduction of errors that can be induced by using a single method.

7.3 Topics Extraction

The identification of covered topics or keywords is of particular interest within our analysis model because it enables us to grasp an overview of a document or a conversation, but also provides valuable information when modeling the interaction between voices. We prefer to address topics in terms of voices (see *3.1.2* Bakhtin's Dialogism as a Framework for CSCL) because, in essence, voices are centered on topics and our analysis can be applied both to explicit dialogue with focus on participant interaction, but also to inner dialogue, when reading a document or expressing one's metacognitions. Therefore, topics, seen as key concepts from within the discourse, play a leading role in obtaining a general perspective, but also in observing emerging points of interest or shifts of focus.

Tightly connected to the cohesion graph, topics can be extracted at different levels and from different constituent elements of the analysis (e.g., the entire document or conversation, a paragraph or all the interventions of a participant). The relevance of each concept mentioned in the discussion and represented by its lemma is determined by combining a multitude of factors:

- *Individual normalized term frequency* – $1 + \log(no\ occurences)$ (Manning et al., 2008); in the end, we opted for eliminating inverse document frequency, as this factor is related to the training corpora and we wanted to grasp the specificity of each analyzed text.
- *Semantic similarities* through the cohesion function (LSA cosine similarity and inverse of LDA Jensen-Shannon divergence) with the analysis element and to the whole document for ensuring global resemblance and significance.
- A *weighted similarity* with the corresponding *semantic chain* multiplied by the importance of the chain; semantic chains are obtained by merging lexical chains determined from the disambiguation graph modeled through semantic distances from *WordNet* and *WOLF* (Galley & McKeown, 2003) through LSA and LDA semantic similarities and each chain's importance is computed as its normalized length multiplied with the cohesion function between the chain, seen as an entity integrating all semantically related concepts, and the entire document.

In this context, key topics together with their corresponding semantic chains can be considered voices that spread throughout the discourse (see *4.2* Discourse Analysis and the Polyphonic Model), while cohesion simulates echoes of voices to other inter-linked textual elements. Moreover, by changing

the focus on a specific participant in a chat conversation, we can observe the strength of a voice as being directly proportional to the relevance of previously identified topics. In addition, as an empirical improvement and as the previous list of topics is already pre-categorized by corresponding parts of speech, the selection of only nouns provided more accurate results in most cases due to the fact that nouns tend to better grasp the conceptualization of the document or of the discussion.

In terms of a document's visualization, the initial text is split into paragraphs, cohesion measures are displayed in-between adjacent blocks and the list of sorted topics with their corresponding relevance scores is presented to the user, allowing him to filter the displayed results by number and by corresponding part of speech. As an example, Figure 42 depicts an excerpt from Williams (2002) presented to 1st year master students during the Natural Language Processing Course 2011-2012.

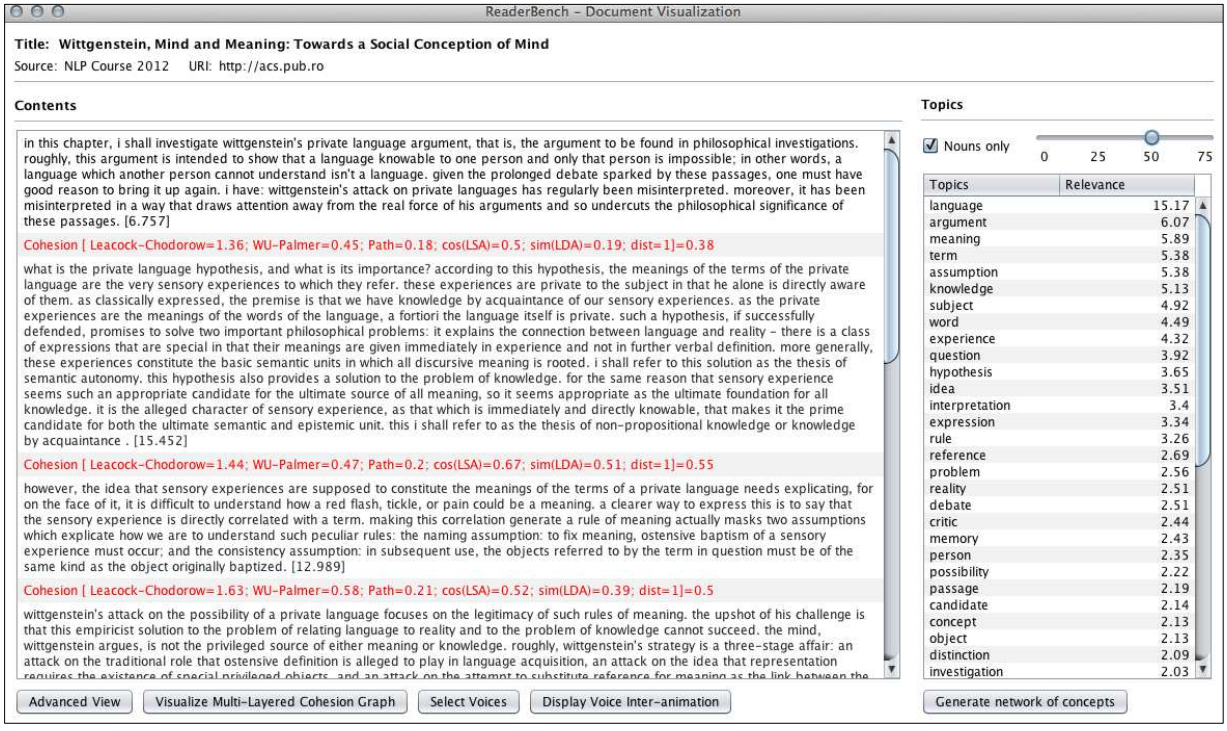


Figure 42. *ReaderBench* (1) Main interface for visualizing documents and topics.

A very interesting extension to topics identification is the visualization of the corresponding semantic space that can also be enlarged with semantically similar concepts, not mentioned within the discourse and referred to in our analysis as *inferred concepts*. Therefore, an inferred concept does not appear in the document or in the conversation, but is semantically related to it. From a computational perspective, the list of additional inferred concepts identified by *ReaderBench* is obtained in two steps. The first stage consists of merging lists of similar concepts for each topic, determined through synonymy and hypernymy relations from *WordNet*/*WOLF* and through

semantic similarity in terms of LSA and LDA, while considering the entire semantic spaces. Secondly, all the concepts from the merged list are evaluated based on the following criteria: semantic relatedness with the list of identified topics and with the analysis element, plus a shorter path to the ontology root for emphasizing more general concepts.

The overall generated network of concepts, including both topics from the initial discourse and inferred concepts, takes into consideration the aggregated cohesion measure between concepts (LSA and LDA similarities above a predefined threshold) and, in the end, displays only the dominant connected graph of related concepts (outliers or unrelated concepts that do not satisfy the cohesion threshold specified within the user interface are disregarded). The visualization uses a Force Atlas layout from *Gephi* (Bastian et al., 2009) and the dimension of each concept is proportional with its betweenness score (Brandes, 2001) from the generated network (see Figure 43).

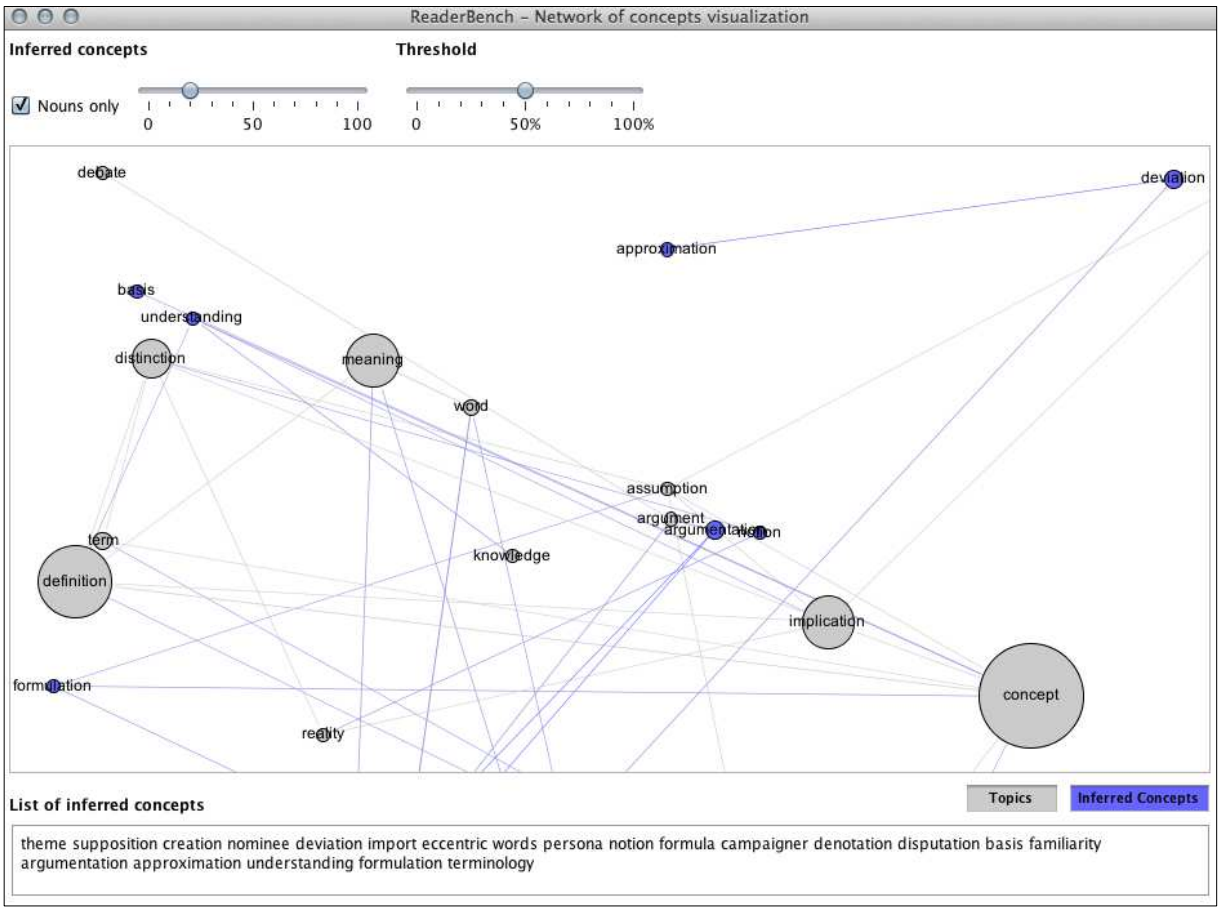


Figure 43. *ReaderBench* (1) Network of concepts visualization from and inferred from Williams (2002).

Although the majority of displayed concepts make perfect sense and really seem close to the given initial text, in most cases there are also some dissonant words that appear to be off-topic at a first glimpse. In the example presented in Figure 43, "*campaigner*" might induce such an effect, but its

occurrence in the list of inferred concepts is determined by its synonymy relationship from *WordNet* to "*candidate*", a concept twice encountered in the initial text fragment that has a final relevance of 2.14 (see Figure 42). Moreover, the concept has only 7 occurrences in the TASA training corpus for LSA and LDA, therefore increasing the chance of making incorrect associations in the semantic models as no clear co-occurrence pattern can emerge.

In this context, additional improvements must be made to the previous identification method in order to reduce the fluctuations of the generated inferred concepts, frequent if the initial topics list is quite limited or the initial text is rather small, and to diminish the number of irrelevant generated terms, by enforcing additional filters. Currently, the identification of inferred concepts was not subject to a formal validation due to the noise detected in smaller text fragments, but all the previously proposed mechanisms were fine-tuned after detailed analyses on different evaluation scenarios and on different types of texts (stories, assigned reading materials and chat conversations), generating in the end an extensible and comprehensive method of extracting topics and inferred concepts.

7.4 Cohesion-based Scoring Mechanism

A central component in the evaluation process of each sentence's importance, of participant's involvement and of collaboration throughout a conversation is our bottom-up intervention scoring method. Although tightly related to the cohesion graph (Dascalu, Dessus, et al., in press) (see *7.2* Cohesion-based Discourse Analysis) that is browsed from bottom to top and is used for augmenting the importance of the analysis elements, the initial individual assessment of each element is based on its topics coverage and their corresponding relevance (see *7.3* Topics Extraction), with respects to the entire document (applicable for both general texts and chat conversations). Therefore, topics are used to reflect the local importance of each analysis element, whereas cohesive links are used to transpose the local impact upon other inter-linked elements.

In terms of the scoring model, each sentence is initially assigned an individual score equal to the normalized term frequency of each concept, multiplied by its relevance that is assigned globally during the topics identification process (see *7.3* Topics Extraction). In other words, we measure to what extent each sentence conveys the main concepts of the overall conversation, as an estimation of on-topic relevance. Afterwards, at block level (utterance or paragraph), individual sentence scores are

weighted by cohesion measures and summed up in order to define the inner-block score. This process takes into consideration the sentences' individual scores, the hierarchical links reflected in the cohesions between each sentence and its corresponding block and all inner-block cohesive links between sentences (internal links within a block that can be defined explicitly, based on adjacency or be automatically identified as relevant cohesive links in the cohesion graph) (see Figure 40).

By going further into our discourse decomposition model (document > block > sentence), inter-block cohesive links are used to augment the previous inner-block scores, by also considering all block-document similarities as a weighting factor of block importance. Moreover, as it would have been a discrepancy in the evaluation in terms of the first and the last sentence of each block for which there were no previous or next adjacency links within the current block, their corresponding scores are increased through the cohesive link enforced to the previous, respectively next block. This augmentation of individual sentence scores is later on reflected in our bottom-up approach all the way to the document level in order to maintain an overall consistency, as each higher level analysis element score should be equal to a weighted sum of constituent element scores.

In the end, all block scores are combined at document level by using the block-document hierarchical link's cohesion as weight, in order to determine the overall score of the reading material or of the conversation. In this manner, all links from the cohesion graph are used in an analogous manner for reflecting the importance of analysis element; in other words, from a computational perspective, hierarchical links are considered weights and are characterized as a spread of information into subsequent analysis elements, whereas adjacency or relevant links between elements of the same level of the analysis are used to augment their local importance through cohesion to all inter-linked sentences or blocks.

Figure 44 presents the main user interface of *ReaderBench* in which a chat conversation has been loaded from the XML format obtained after the conversion of the HTML file exported from *ConcertChat* (Holmer et al., 2006) according to the schema designed in *Polyphony* (Trausan-Matu, Rebedea, et al., 2007) (see Appendix D – Input Examples, Sample Chat – Log of Team 4 Chat Conversation). The importance scores of each utterance are displayed in brackets, whereas the automatically identified topics for a specific participant through a process equivalent to the one described in detail in *7.3* Topics Extraction, but applied on the set of each individual's interventions, are presented in the right sidebar. Additional interaction- and collaboration-centered functionalities

are provided to the user, later to be detailed in *9 ReaderBench* (3) – Involvement and Collaboration Assessment. An equivalent visualization is available also for general texts (see Figure 45).

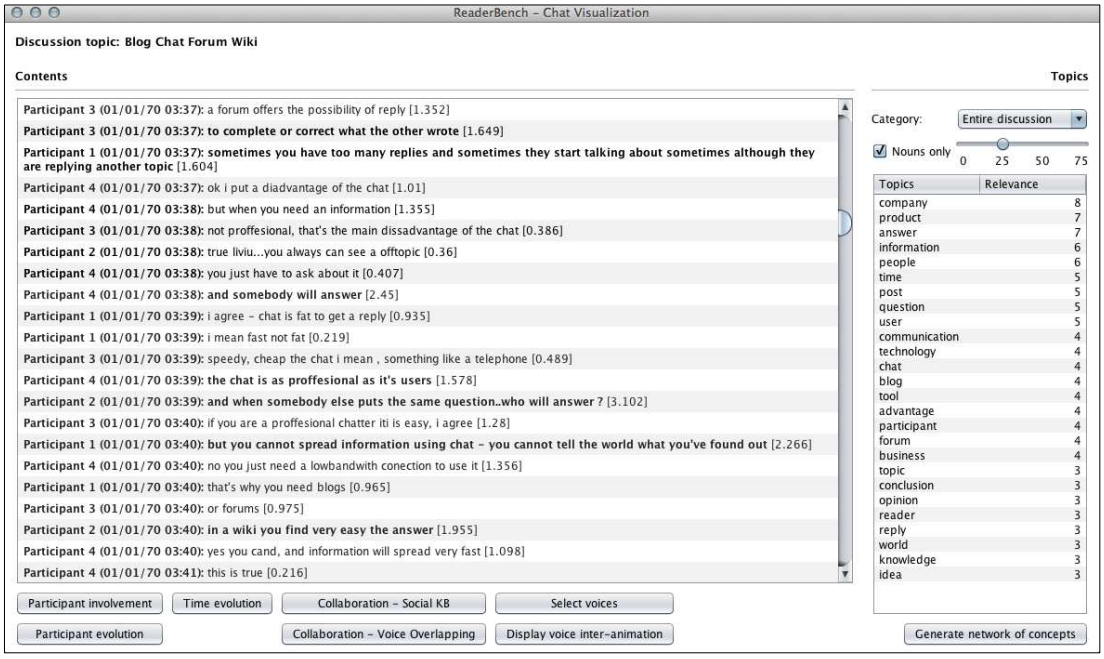


Figure 44. *ReaderBench* (1) Chat conversation visualization.
Including utterance importance scores (in square brackets after each intervention), demarcation with bold of utterances considered most important according to the summarization facility and participant topics

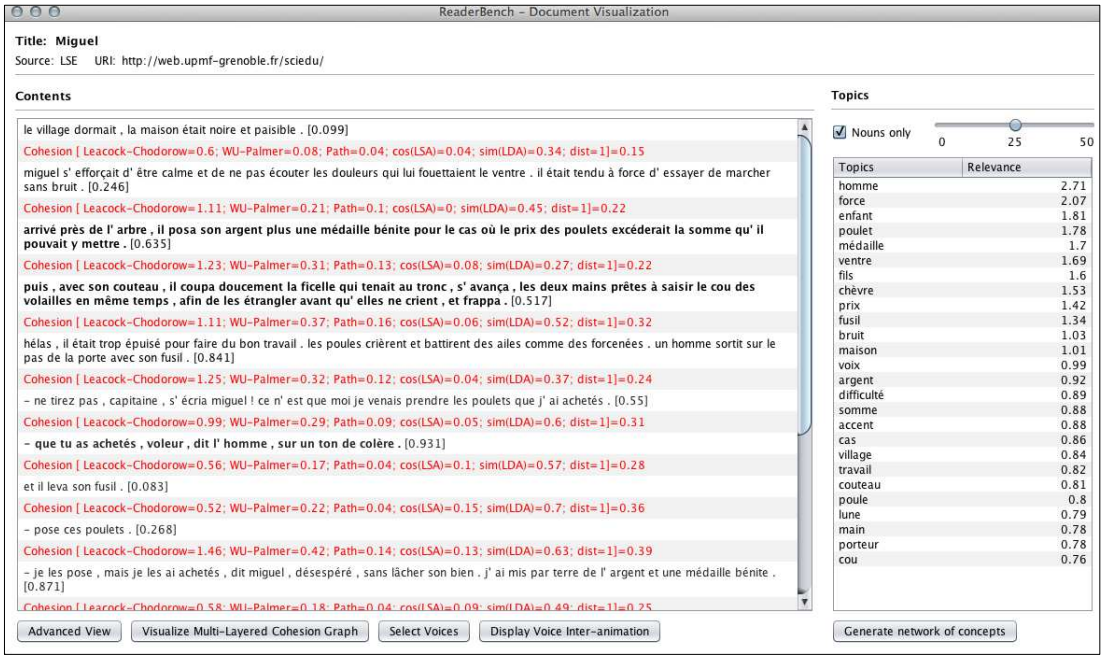


Figure 45. *ReaderBench* (1) Reading material visualization.
Including block scores (in square brackets after each paragraph), demarcation with bold of sentences considered most important according to the summarization facility and document topics. Although the block score can be elevated (e.g., "*hélas* …"), it is a combination of individual sentence scores; therefore, underlying sentences might not be selected in the summarization process

In addition, starting from the tutors' observations during the two validation rounds of *PolyCAFe* (Rebedea et al., 2010; Rebedea, Dascalu, Trausan-Matu, Armitt, et al., 2011) that the extractive summarization facility (see *6.4.3* Semantic Extractive Summarization and *6.5* Validation of *PolyCAFe*), combined with the demarcation of the most important utterances, was useful for providing a quick overview or a recap of the conversation, we envisioned an *extractive summarization* facility within *ReaderBench*. This functionality can be considered a generalization of the previous scoring mechanism built on top of the cohesion graph and can be easily achieved by considering the sentence importance scores, in descending order, as we are enforcing a deep discourse structure, topics coverage and the cohesive links between analysis elements. Overall, the proposed unsupervised extraction method is similar to some extent to *TextRank* (Mihalcea & Tarau, 2004) that also used an underlying graph structure based on the similarities between sentences. Nevertheless, our approach can be considered more elaborate from two perspectives: 1/ instead of simple word co-occurrences we use a generalized cohesion function (see *7.2* Cohesion-based Discourse Analysis) and 2/ instead of computing all similarities between all pairs of sentences, resulting in highly connected graph, inapplicable for large text, we propose a multi-layered graph that resembles the core structure of the initial texts in terms of blocks or paragraphs.

Although this summarization facility can be applied on both conversations and general texts, as *preliminary validation* we have performed experiments on two narrative texts in French: "*Miguel de la faim*" (Vidal, 1984) (see Figure 45) and "*La pharmacie des éléphants*" (Pfeffer, 1989), starting from the measurements initially performed by Mandin (2009). Moreover, due to the high discrepancy and variance in the actual dimension of utterances and of the overall conversation, a relevant validation scenario in terms of CSCL conversations was rather difficult to achieve because: 1/ the selection criteria differed greatly between annotators, although the general tendency when resuming chat conversation was to select longer and on-topic interventions and 2/ selecting 10% or 20% of the conversations seemed rather difficult to grasp by evaluators, whereas selecting the most important 20 interventions induced too much noise in the final evaluation; nevertheless, setting an imposed number of utterances would have been unfeasible due to high discrepancies in conversation length. Therefore, we focused our analysis on general texts for high school (9^{th}-12^{th} grade) students and tutors were asked to manually highlight the most important 3 to 5 sentences from the two presented stories (see Table 21 for general statistics) (Mandin, 2009).

Table 21. *ReaderBench* (1) Summarization evaluation statistics.

Text	No. sentences	Number of evaluators per level					Intraclass Correlation Coefficient	
		9th grade	10th grade	11th grade	12th grade	Tutor	Single	Average
Miguel de la faim	24	30	39	68	19	9	.13	.96
La pharmacie des éléphants	18	28	39	85	22	16	.23	.98

Correlation results between the sentence scores and the average rankings per evaluator level are presented in Table 22. Although the average correlation of the scoring mechanism is quite low, its corresponding value is still better than the reliability of a single rater.

Table 22. *ReaderBench* (1) Correlation between automatic sentence scores and manual rankings.

Text	Correlation to average evaluator ranking					Average correlation
	9th grade	10th grade	11th grade	12th grade	Tutor	
Miguel de la faim	.37	.14	.39	.29	.26	.29
La pharmacie des éléphants	.28	.38	.18	.31	.17	.26

Afterwards, as suggested by Donaway, Drummey, and Mather (2000) as a simple binary categorization of importance is in most cases insufficient, four equivalence classes were defined, taking into consideration the $mean - stdev$, $mean$ and $mean + stdev$ of each distribution as cut-out values. In this context, two measurements of agreement were used: *exact agreement* (EA) that reflects precision and *adjacent agreement* (AA) that allows a difference of one between the class index automatically retrieved and the one evaluated by the human raters. By considering the use of the equivalence classes, we notice major improvements in our evaluation (see Table 23) as both documents have the best agreements with the tutors, suggesting that our cohesion-based scoring process entails a deeper perspective of the discourse structure, reflected in each sentence's importance.

Table 23. *ReaderBench* (1) Exact and Adjacent Agreement between automatic and manual sentence selection using equivalence classes.

Text	Exact/Adjacent Agreement (EA/AA)					Average EA/AA
	9th grade	10th grade	11th grade	12th grade	Tutor	
Miguel de la faim	.33/.83	.42/.75	.29/.88	.38/.88	.46/.88	.38/.84
La pharmacie des éléphants	.22/.83	.28/.89	.33/.78	.39/.94	.44/.89	.33/.87

Moreover, our results became more cognitively relevant as they are easier to interpret by both learners and tutors – instead of a positive value obtained after applying the scoring mechanism, each sentence has an assigned importance class (1 – less important; 4 – the most important). In addition, we obtained 3 or 4 sentences per document that were tagged with the 4th class, a result consistent with the initial annotation task of selecting the 3-5 most important sentences. Therefore, based on promising preliminary validation results, we can conclude that the proposed cohesion-based scoring mechanism is adequate and effective, as it integrates through cohesive links the local importance of each sentence, derived from topics coverage, into a global view of the discourse.

7.5 Dialogism and Voice Inter-Animation

The key element in terms of voice identification (see *3.1.2* Bakhtin's Dialogism as a Framework for CSCL) resides in building lexical chains and merging them into semantic chains through cohesion. Due to the limitation of discovering lexical chains (Galley & McKeown, 2003) through semantic distances in *WordNet* (Miller, 1995) or *WOLF* (Sagot, 2008) that only consider words having the same part-of-speech (see *4.3.1* Semantic Distances and Lexical Chains), the merge step is essential as it enables consideration of different parts-of-speech and unites groups of concepts based on identical lemmas or high cohesion values. In this context, we have proposed an iterative algorithm similar to an agglomerative hierarchical clustering algorithm (Hastie, Tibshirani, & Friedman, 2009) that starts with the identified lexical chains seen as groups of already clustered words and uses as distance function the cohesion between the corresponding groups of words, if this value is greater than an imposed threshold, in order to merge clusters.

As semantic chains span across the discourse, the context generated by the co-occurrence or repetitions of tightly cohesive concepts is similar to the longitudinal dimension of voices. Echoes can be highlighted through cohesion based on semantic relationship and attenuation is reflected in the considered distance between analysis elements (see *7.2* Cohesion-based Discourse Analysis). Moreover, by intertwining different semantic chains within the same textual fragment (sentence, utterance or paragraph) we are able to better grasp the transversal dimension of voice inter-animation. Therefore, after manually selecting the voices of interest, the user can visualize the conversation as an overlap of co-occurring semantic chains that induce polyphony (see Figure 46). A voice is displayed within the interface as the 3 most dominant concepts (word lemmas) and its occurrences throughout the conversation are marked accordingly to the overall timeframe. Different

speakers that uttered a particular voice are demarcated with randomly assigned colors, consistent throughout a conversation for each participant. Each utterance may incorporate more than a single voice, as it may include, in addition to the current participant's voice, at least one other, an alien voice (Bakhtin, 1981; Trausan-Matu & Stahl, 2007) (see *3.1.2* Bakhtin's Dialogism as a Framework for CSCL), identified through semantic chains and cohesive links.

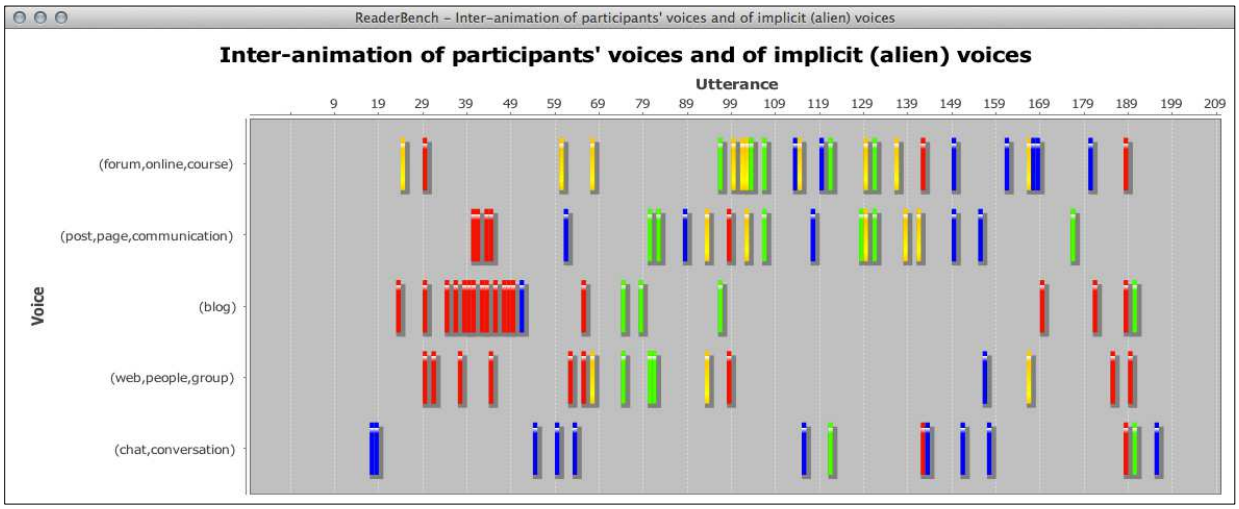


Figure 46. *ReaderBench* (1) Chat voice inter-animation visualization covering participants' voices and implicit (alien) voices.
The chart follows the conversation timeline expressed in utterance identifiers and depicts the occurrences of the 5 most dominant voices: 1/ (forum, online, course), 2/ (post, page, communication), 3/ (blog), 4/ (web, people, group) and 5/ (chat, conversation). Each of the 4 chat participants has a corresponding color and each voice occurrence reflects the speaker's assigned color

In order to better grasp the importance of each voice within the discourse, we have devised a series of indicators, some inspired from 'rhythmanalysis' (Lefebvre, 2004) and 'polyrhythm' (*The New Harvard Dictionary of Music*, 1986): 1/ the number of contained words as a pure quantitative factor, 2/ the cumulative scores of the analysis elements that provides a broader image of the importance of the context of their occurrence (qualitative oriented) and 3/ the recurrence of voices seen as the distance between two analysis elements in which consecutive occurrences of the voice appear, inspired from rhythm analysis.

Moreover, in accordance to Miller's law (Miller, 1956), we have applied a simple moving average (Upton & Cook, 2008) on the voice distribution for five datum points representing consecutive utterances (or sentences in the case of general texts), with a split horizon of one minute between adjacent interventions (only for chat-based conversations where the timestamp of each utterance is used). In other words, we weight the importance of each concept occurrence over 5 adjacent

utterances, if no break in the discourse is larger than an imposed, experimentally determined threshold of one minute. Exceeding this value would clearly mark a stopping point in the overall chat conversation, making unnecessary the expansion of the singular occurrence of the voice over this break. This step of smoothing the initial discrete voice distribution plays a central role in subsequent processing as the expanded context of a voice's occurrence is much more significant than the sole consideration of the concept uttered by a participant in a given intervention. In this particular case, entropy (Shannon, 1948) has been applied on the smoothed distribution in order to highlight irregularities of voice occurrences throughout the entire conversation.

By considering all the previous factors used to estimate the importance of a voice, Table 24 presents an image of their correlations when considering a conversation of approximately 400 interventions and all 57 automatically identified voices, with the sole constraint that each voice had to include at least 3 word occurrences in order to have a quantifiable overall impact. Overall, all factors, besides recurrence, correlate positively and can be used to estimate the overall impact of a voice within the conversation, whereas recurrence is more specific and can be used to pinpoint whether the concepts pertaining to a voice are collocated or are more equally dispersed throughout the discourse. Nevertheless, small correlation values are acceptable as our aim was to identify meaningful factors that can be used to better characterize a voice's importance. Further evaluations need to be performed in order to determine the most representative factors, but our aim was to identify specific measures of evaluation that are generated as effects of different underlying assessment factors (e.g., the use of the number of utterances in which the voices occurred or of statistics applied on the initial distribution would have been inappropriate as all these factors would have been directly linked to the number of words within the semantic chain).

Table 24. *ReaderBench* (1) Cross-correlations matrix for voice analysis factors.

	1	2	3	4	5
1. Number of words within the semantic chain	1	.20	.77	-.44	-.35
2. Average utterance importance scores	.20	1	.26	-.20	-.08
3. Entropy applied on the utterance moving average	.77	.26	1	-.68	-.46
4. Recurrence Average	-.44	-.20	-.68	1	.67
5. Recurrence standard deviation	-.35	-.08	-.46	0.67	1

As voice synergy emerges as a measure of co-occurrence of semantic chains, mutual information (Manning et al., 2008) can be used to quantify the global effect of voice overlapping between any pairs of voices (see Figure 47). Moreover, by applying pointwise mutual information (PMI) (Fano, 1961) between the moving averages of all pairs of voice distributions that appear in a given context of five analysis elements, we obtain a local degree of voice inter-weaving or overlap. In order to better grasp the underlying reason of using PMI, we have presented in Figure 48 three progressive measures for synergy.

Correlation Matrix

	0	1	2	3	4
0 - (forum,online,course)					
1 - (post,page,communication)	5.166				
2 - (blog)	5.928	5.244			
3 - (web,people,group)	5.41	5.095	4.831		
4 - (chat,conversation)	5.437	5.618	6.043	5.319	

Figure 47. *ReaderBench* (1) Mutual information between pairs of voices (cross-correlations matrix).

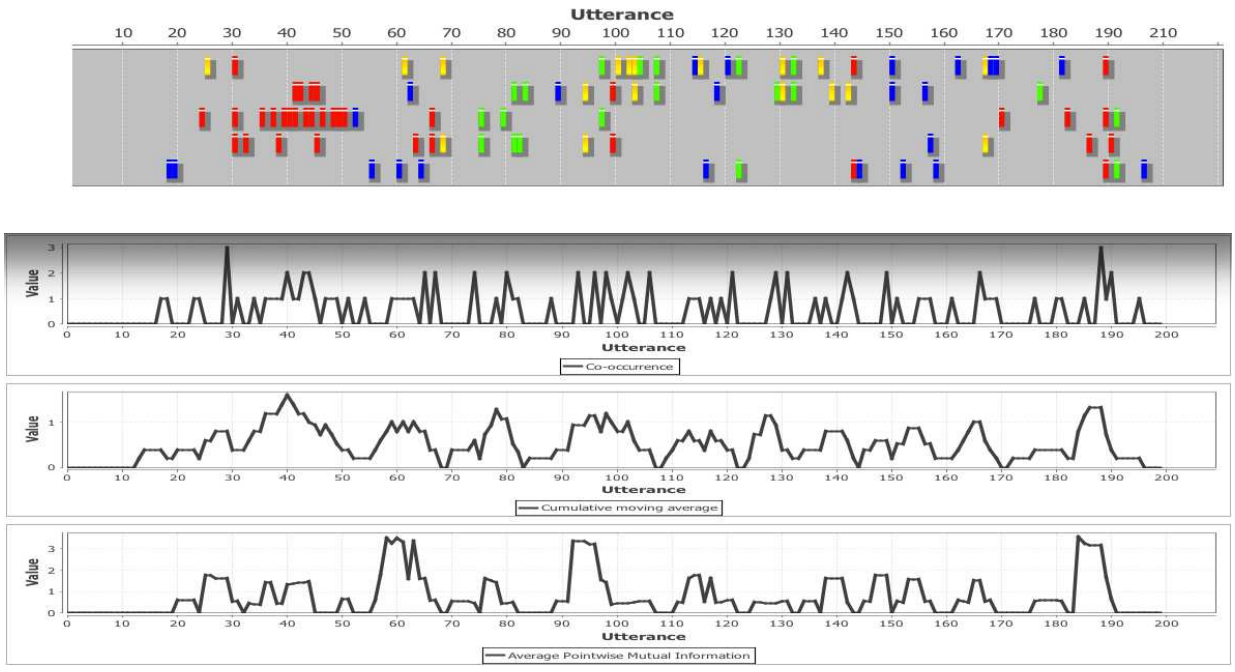


Figure 48. *ReaderBench* (1) Evolution of voice synergy throughout the conversation.
a. Voice visualization as time evolution (baseline for comparison); b. Number of occurrences; c. Evolution of cumulated moving average; c. Average pointwise mutual information

The *first* and the simplest, the actual number of voices (co-)occurring, is misleading as we encounter a lot of singular values (meaningless as only one voice is present) and double ones, which are also not that interesting in observing the global trend. Also, the first spike with a value of 3 is locally representative, but since it's isolated from the rest of the conversation, its importance should be mediated globally. The *second*, the cumulated moving average, is better as the smoothing effect has a positive impact on the overall evolution. Nevertheless, it is misleading in some cases – e.g., the maximum value is obtained around utterance 40 where the conversation is dominated by one participant and one voice, but by being so strong, even the smoothed effect is artificially augmented.

The third, the average PMI applied on the moving averages, grasps best the synergic zones: e.g., just after utterance 60 we have all five selected voices co-occurring, between 95 and 100 an overlap of four voices, the first two being well represented and dominant, and just before utterance 190 we also have four co-occurring voices. Therefore, by observing the evolution of PMI using a sliding window that follows the conversation flow, we obtain a trend in terms of synergy that can be later on generalized to Bakhtin's polyphony (Bakhtin, 1984).

We opted for presenting the evolution of voice synergy instead of polyphony because our computational model uses co-occurrences and overlaps of voices within a given context. In order to emphasize the effect of inter-animation that would induce true polyphony, we envisage the use of argumentation acts and patterns (Stent & Allen, 2000) for highlighting the interdependencies between voices and how a particular voice can shed light on another.

Starting from the common components presented in this chapter addressing general documents and focusing on discourse analysis (cohesion-based or voice-based), topics extraction and the scoring mechanism, the following chapters will be centered on: 1/ *individual assessment* through the identification of reading strategies and the evaluation of textual complexity (see *8 ReaderBench* (2) – Individual Assessment through Reading Strategies and Textual Complexity) and 2/ *involvement and collaboration assessment* through the use of the cohesion graph, of utterance importance scores, and of the two collaboration assessment models based on social knowledge-building and voice inter-animation (see *9 ReaderBench* (3) – Involvement and Collaboration Assessment through Cohesion and Dialogism).

8 *ReaderBench* (2) – Individual Assessment through Reading Strategies and Textual Complexity

As an overview, in terms of individual learning, *ReaderBench* encompasses the functionalities of both *CohMetrix* (McNamara et al., 2010) (see *2.2.2* Textual Complexity Computational Approaches) and *iStart* (Graesser, McNamara, & VanLehn, 2005; McNamara, Boonthum, et al., 2007) (see *2.3* Reading Strategies), as it provides teachers and learners information on their reading/writing activities: initial textual complexity assessment, assignment of texts to learners, capture of meta-cognitions reflected in one's textual verbalizations, and reading strategies assessment (a detailed comparison is presented at the end of this chapter). Moreover, *ReaderBench* encompasses textual complexity measures similar to *Dmesure* (T. François, 2012; T. François & Miltsakaki, 2012), but with emphasis on more in-depth, semantic factors. The main differentiators between *ReaderBench* and the previous systems consist of the following (see *8.3* Comparison of *ReaderBench* to *iSTART, Dmesure and Coh-Metrix* for more details):

- Emphasis on comprehension extracted from the automatic analysis of metacognitions (Dascalu, Dessus, et al., in press), based on two preliminary studies (Dessus et al., 2012; Oprescu et al., in press).
- A different educational purpose, as *ReaderBench* validation was performed on primary school pupils, whereas *iStart* mainly targets high school and university students (Nardy et al., in press).
- Different factors, measurements and the use of SVMs (Cortes & Vapnik, 1995; T. François & Miltsakaki, 2012) for increasing the validity of textual complexity assessment (Dascalu et al., 2012).

8.1 Identification of Reading Strategies

The use of reading strategies is widely recognized as a crucial determinant of reading comprehension (see *2.3* Reading Strategies). Second degree and high school pupils who are good comprehenders are mostly strategic readers (Graesser, 2007). These strategies can be elicited through self-explanations (Chi et al., 1994) and have been categorized by McNamara (2004) as follows: comprehension monitoring, paraphrasing, elaboration, prediction, and bridging. One important skill that these strategies exploit is to be able to establish semantic and causal relationships between the read sentences (Wolfe, Magliano, & Larsen, 2005).

Based on these findings, McNamara, Boonthum, et al. (2007) developed *iSTART*, a cognitive tutor that automatically categorizes self-explanations, partly using Latent Semantic Analysis (Landauer & Dumais, 1997). Any thorough analysis of self-explanations reports it is a very demanding and subjectivity-oriented activity, and the use of systems like *iSTART* to detect pupils' reading strategies is more than challenging. Since a cognitive tutor guides the reader through pre-defined steps alternating between reading and verbalizations, a similar computer-based scenarization is made possible through the wide range of reading strategies and the feedback possibilities (Vitale & Romance, 2007). Nevertheless, as our focus was to automatically assess verbalizations and to identify reading strategies, multiple alternatives were explored: two initial studies addressed in extent the identification of paraphrases (Dessus et al., 2012; Oprescu et al., in press), while an integrated view targeting the automatic identification of all proposed reading strategies (both low-level – causality, control, paraphrasing – and high-level, cognitive strategies – knowledge inference and bridging) was first introduced in *ReaderBench* (Dascalu, Dessus, et al., in press).

The *data gathering* and *evaluation method* applied a priori was the same for all experiments, but the corpus of evaluated verbalizations consisted of different sub-sets of the entire collection. In the end, during the ANR DEVCOMP project, 84 pupils from 3rd to 5th grades, from the same school and from a middle socio-economic background participated in our experiments. The pupils read a narrative text consisting of 453 words, the story "*Matilda*" by Dahl (2007), and explained what they understood up to that point at 6 predefined breakpoints (see Appendix D – Input Examples, Sample Document – Matilda by Dahl (2007) for complete text). The text was chosen to be within the reading level of participants, so that differences in verbalizations would indicate differences in reading strategies instead of comprehension difficulties. In order to perform a fine-grained analysis, the initial

text was split in 45 segments (of about 1 sentence each). A causal analysis was performed so that both local (when the causal antecedent is close to the reference sentence) and distal antecedents (when the causal antecedent is somewhat farther, out of the reader's working memory) of sentences were determined in accordance to Millis, Magliano, and Todaro (2006). Finally, a propositional analysis of the text was proposed that allowed us to extract macro-propositions and to support the coding of what was remembered by the participants.

Participants individually read the text out loud and stopped at predetermined breaks to self-explain the text segment just read, the whole activity being recorded. The task was explained to pupils as follows: "During your reading you will stop at each icon to tell out loud what you have understood, just at this time". Their verbalizations were then transcribed and each self-explanation was semantically compared using different natural language processing techniques. Pupils' verbalizations were analyzed proposition by proposition and were categorized by experts according to a coding scheme adapted from McNamara (2004). Disagreements between experts in terms of identified reading strategies were discussed and resolved by consensus (Nardy et al., in press). As technical specificity, the first two studies were conducted using LSA vector spaces trained on the "*TextEnfants*" corpus (Denhière et al., 2007) (approx. 4.2 M words) with no specific NLP or Information Retrieval optimizations (only stop words elimination), while *ReaderBench* also integrated "*Le Monde*" corpus (French newspaper, approx. 24 M words) with all optimizations mentioned in *7.2* Cohesion-based Discourse Analysis.

8.1.1 The Initial Study of Analyzing Paraphrases

The first study (Dessus et al., 2012) focused on how two main kinds of sentences are paraphrased: *focal* (the latest sentence before a verbalization) and *causal* sentences (identified by a hand-made causal analysis of the text), because it was worth distinguishing the mere paraphrase of the latest read sentence and more elaborated paraphrases, involving a deeper comprehension of the read text. For this experiment, we used a subset of the aforementioned participants sample, consisting of 22 third and 22 fifth grade pupils. Moreover, this study does not involve *ReaderBench*, but it provided a strong experimental base in terms of analyzing paraphrases.

Our research questions were: 1/ to compare human expert categorization of paraphrases to the semantic similarity between text sentences and self-explanations, obtained by means of LSA; 2/ to

highlight an expected "recency effect", stating that the information children self-explain most often pertains to very close sentences to the verbalization break; 3/ to investigate the way pupils account for causal relations (either local or distal) in retelling causally related text sentences.

Firstly, we computed accuracy measures in order to compare human vs. LSA values of sentence relatedness and to check the validity of the computer-based measures. Pearson correlations between the number of paraphrases per verbalization (V_n) detected by the two raters and LSA similarities between each verbalization and the previous sentences were as follows: V_1: $r = .48$; V_2: $r = .58$; V_3: $r = .74$; V_4: $r = .29$; V_5: $r = .57$; V_6: $r = .61$, which shows that human judgments of paraphrases expressed by children on each paragraph are moderately to strongly related to LSA measures of similarities.

Secondly, we investigated the extent to which each self-explanation was related to the last read sentence (focal) (see Figure 49). We observed that the recency effect varies across verbalization plots, indicating that this effect is dependent of the content conveyed by the last sentences. Moreover, the focal sentence, in general, does not have a higher similarity with the related verbalization than the average of other previous sentences, except for V_4: $t(43) = 7.5$, $p < .0005$. Two-way ANOVAs showed a significant difference between grades for V_6, $F(1, 42) = 7.01$; $p < .05$ and a tendency for V_2, $F(1, 42) = 3.22$, $p < .09$. Although grade 3 pupils tended to recall the last sentence at these points more frequently, the semantic content of the last sentence seems to be the main determinant of focal recall.

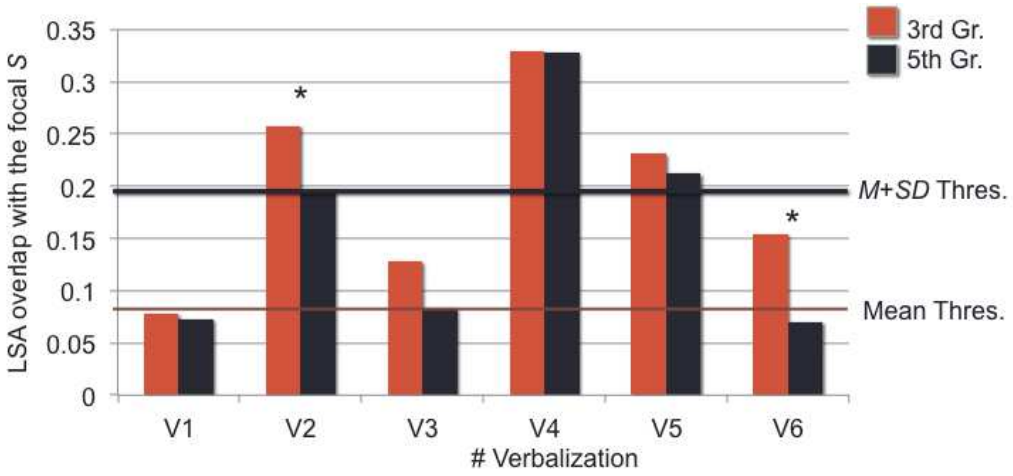


Figure 49. *Matilda* – Mean LSA-based values for similarity of focal sentences by grade.

Thirdly, we initially expected that 1/ the semantic content of local and distal sentences, as determined by the causal analysis, is more often verbalized than the rest of the previous text and the

focal sentence and 2/ the local-centered causal sentences are better recalled than the distal-centered ones (see Figure 50). Results first showed that local and distal causal sentences are, in all cases but two (local vs. V_1 and V_5), significantly more verbalized than the rest of the text. Moreover, the content of local causal sentences was significantly better recalled than focal sentences in V_1 and V_3 (resp. $t(43) = 3.11$, $p < .005$; $t(43) = 9.45$, $p < .0005$). Unexpectedly, the content of distal causal sentences was better recalled than local causal sentences for V_1: $t(43)=6.09$, $p < .0005$; V_2: $t(43)=8.49$, $p < .0005$. Two-way ANOVAs showed significant differences between grades for V_1 (distal), $F(1, 42) = 4.43$, $p < .05$; and a tendency for V_6 (distal), $F(1, 42) = 3.90$, $p < .06$ and for V_3 (local), $F(1, 42) = 2.91$; $p < .1$. Overall, participants' strategies focused on causality, rather than recency.

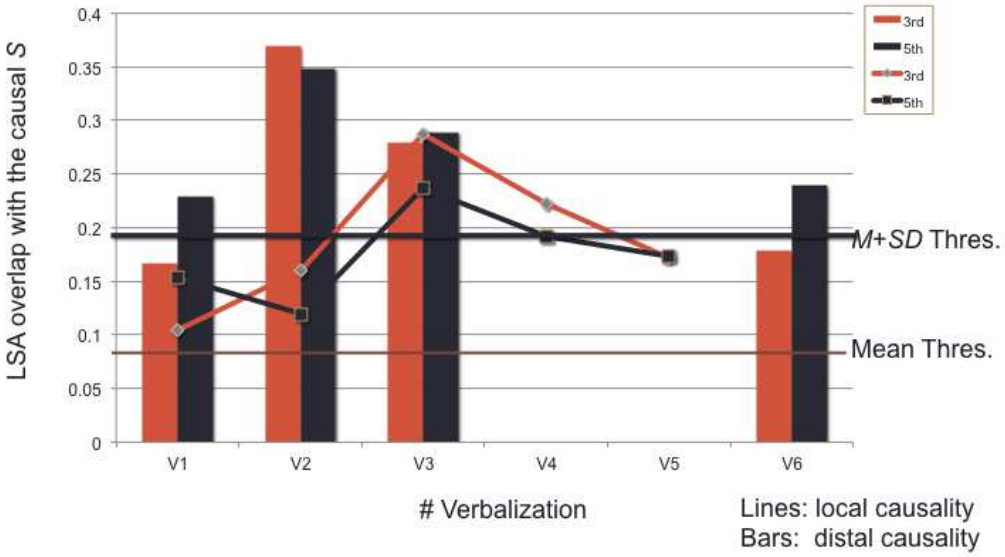


Figure 50. *Matilda* – Mean LSA-based values for similarity of causal sentences, by grade. Lines: local causality; bars: distal causality.

In conclusion, the initial study presented a first attempt to set up the foundations of a cognitive reading tutor aiming at analyzing pupils' verbalizations to get some traces of their strategies. The results showed that LSA-based analyses of verbalizations correlate moderately to high with those of human experts and therefore founding our analysis on LSA derived metrics is meaningful. Additionally, and as also shown by Trabasso and van den Broek (1985), participants tended to recall sentences they read according to causality-driven, rather than recency-driven strategies, which reveal to some extent their comprehension strategies. Eventually, there was also a grade effect on the way distal and local causal sentences are recalled that required further investigations.

8.1.2 The Second Study of Analyzing Paraphrases

The second study (Oprescu et al., 2012; Oprescu et al., in press) focused on evaluating paraphrases by enforcing different natural processing techniques and by comparing two heuristics – *word-based* and *LSA similarity* – in order to establish further research paths. For implementing the *word-based heuristic*, *Tree Tagger* (H. Schmidt, 1994, 1995) and *WOLF* (Sagot, 2008; Sagot & Darja, 2008) are used for creating lists of relevant words, classified by corresponding part of speech, for each paragraph and verbalization. Then the fraction between the words in the paragraph and the words in the verbalization is computed for each category by considering also synonymy relations from WOLF. Four fractions are obtained and a weighted average of the four is returned as an overall rating (see Equation 27).

$$R_W = \frac{W_n \frac{n_n}{N_n} + W_v \frac{n_v}{N_v} + W_{aj} \frac{n_{aj}}{N_{aj}} + W_{av} \frac{n_{av}}{P_{av}}}{W_n + W_v + W_{aj} + W_{av}} \quad (27)$$

where R_W is the rating returned by the function, n_n, n_v, n_{aj} and n_{av} are the number of nouns, verbs, adjectives and respectively adverbs in the verbalization that can be found in the list of relevant nouns of the paragraphs, N_n, N_v, N_{aj} and N_{av} are the length of these lists and, and W_n, W_v, W_{aj} and W_{av} are their weights in the average. All these predefined weights were determined experimentally, after running multiple iterations with incremental values.

The *LSA similarity heuristic* compares each sentence of the paragraph to the entire verbalization and a weighted average of the values is computed, ignoring the two smallest values due to the fact that each verbalization usually contains one or more control phrases that are irrelevant to the comparison and may alter the results (e.g., "*j'ai compris que*", "*je me rappelle que*"). The weight of an utterance is equal to the number of words it contains. The whole paragraph is also compared to the verbalization, as we know that the meaning of the paragraph as a whole can be slightly different from the meaning of each sentence individually. In this manner we cover both cases when a verbalization focuses on the whole paragraph or only on some sentences within.

At this point we had introduced two metrics, both indicating the degree of resemblance of two paragraphs, but we had to decide whether the results of these two metrics are coherent or not, so we tried to evaluate the correspondence between the two metrics (see Figure 51). Based on these

observations, we decided that the best way to combine these two metrics was to multiply them. The combined metrics is also represented in the same chart.

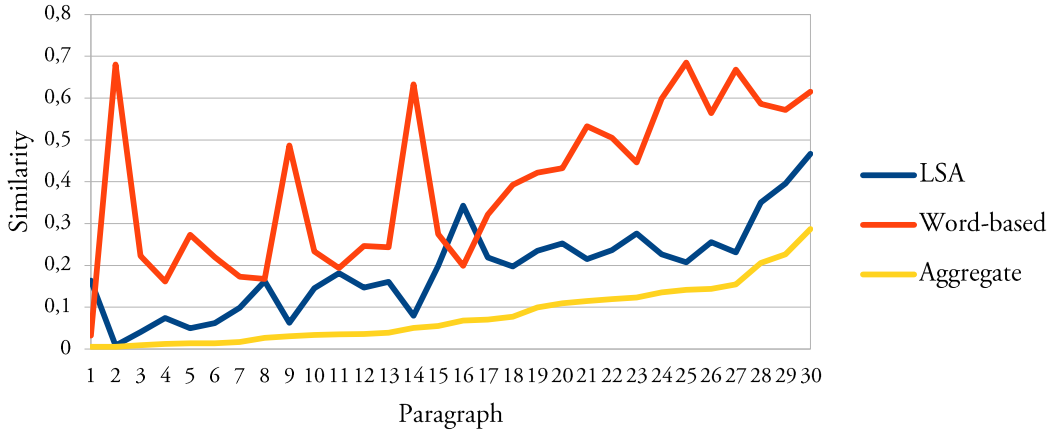


Figure 51. Comparison between the LSA similarity and word-based heuristics.

The Pearson correlation between our metrics was rather low ($r = .34$) since they addressed paraphrasing at lexical and semantic levels, but, as expected, the correlations of each individual heuristic and the aggregated function are much higher ($r_{LSA} = .88$, $r_{word\text{-}based} = .68$); in the end, the LSA metric had a bigger influence on the final similarity score. By observing these results, we decided to establish a threshold for paraphrases around 0.07, determined experimentally. This value allowed us to identify 19 out of the 27 paraphrases identified by human evaluators, which means that we were able to correctly identify 70% of the paraphrases.

Additionally, as a preliminary step to identifying other types of verbalizations, we compared the values of the current paragraph with the previous and the future ones in order to determine the similarity between verbalizations of the same type. As a particularity of this analysis, all initial paragraphs in-between two adjacent verbalizations were merged into a single block of text for better grasping the extent to which different significant text fragments were recalled.

Figure 52 shows the values returned by the word-based metrics for ten paraphrases, which represent about one third of the total number of paraphrases of our test corpus, when compared to the previous, the current and the next segment of text that consists of a merge of all paragraphs in-between two adjacent self-explanations. It is obvious that there is higher resemblance between the current textual segment and the verbalization (so the one just in front of the metacognition break, recalled by the pupil), while the similarity between the verbalization and other surrounding textual segments is close to zero. There are some exceptions, mainly highlighting different types of

verbalizations, but no straightforward conclusion could be drawn and further experiments needed to be conducted.

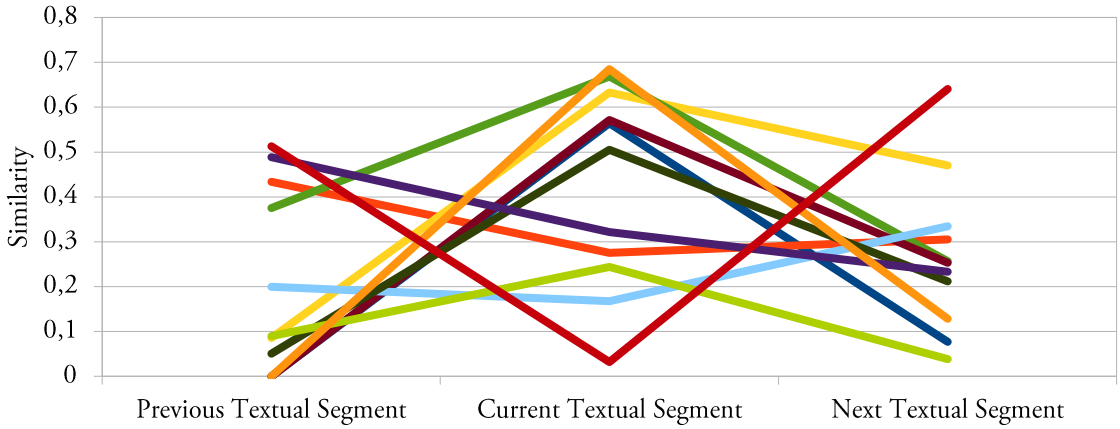


Figure 52. Comparison of verbalizations containing paraphrases, using the word-based heuristic.

Figure 53 depicts a similar analysis using the LSA similarity function. We notice that the graphic has the same characteristics as Figure 52, a similar coefficient of variation (c_v = 0.7), but follows more strictly the pattern of positive slope followed by a negative one, which led us to conclude that the LSA method is more accurate than word-based heuristic, although the average similarity values were quite low. Therefore, in this second study we used LSA and a word-based heuristic to compare the verbalizations with nearby paragraphs and this approach provided encouraging results, as we were able to identify paraphrases with good precision. As conclusions, we decided to focus on extracting reading strategies only by comparing the verbalizations to the previous blocks of texts, in-between the previous and the current verbalization. Moreover, the combination of semantic distances from ontologies and LSA seemed a good practice that lead to the aggregated cohesion function integrated in *ReaderBench*.

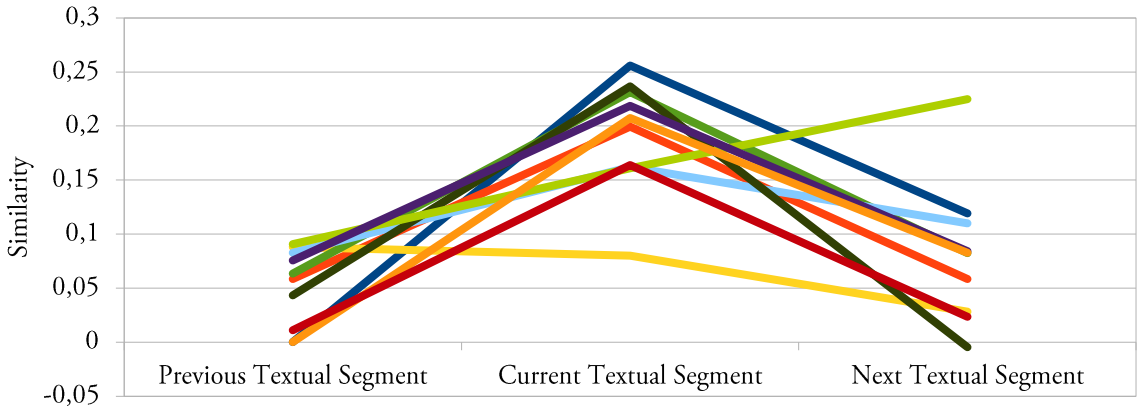


Figure 53. Comparison of verbalizations containing paraphrases, using the LSA-based heuristic.

8.1.3 Reading Strategies Identification Heuristics

Starting from the two previous studies and the five types of reading strategies used by McNamara, O'Reilly, et al. (2007), our aim was to integrate within *ReaderBench* automatic extraction methods designed to support tutors at identifying various strategies and to best fit the aligned annotation categories. The automatically identified strategies within *ReaderBench* comprise monitoring, causality, bridging, paraphrase and elaboration due to 2 observed differences: 1/ very few predictions were used, perhaps due to the age of the pupils, compared to McNamara's subjects; 2/ there is a distinction in *ReaderBench* between causal inferences and bridging, although a causal inference can be considered a kind of bridging, as well as a reference resolution, due to their different computational complexities. Moreover, our objective was to define a fine-grained analysis in which different valences generated by both the identification heuristics and the hand coding rules were taken into consideration when defining the strategies taxonomy. In addition, we have tested various methods of identifying reading strategies and we will focus solely on presenting the alternatives that provided in the end the best overall human-machine correlations.

In ascending order of complexity, the simplest strategies to identify are *causality* (e.g., "*parce que*", "*pour*", "*donc*", "*alors*", "*à cause de*", "*puisque*") and *control* (e.g., "*je me souviens*", "*je crois*", "*j'ai rien compris*", "*ils racontent*") for which cue phrases have been used. Additionally, as *causality* assumes text-based inferences, all occurrences of keywords at the beginning of a verbalization have been discarded, as such a word occurrence can be considered a speech initiating event (e.g., "*Donc*"), rather than creating an inferential link. Afterwards, *paraphrases*, that in the manual annotation were considered repetitions of the same semantic propositions by human raters, were automatically identified through lexical similarities. More specifically, words from the verbalization were considered paraphrases if they had identical lemmas or were synonyms (extracted from the lexicalized ontologies – *WordNet*/*WOLF*) with words from the initial text. In addition, we experimented identifying paraphrases as the overlap between segments of the dependency graph (combined with synonymy relations between homologous elements), but this was inappropriate for French as there is no support within the Stanford Log-linear Part-Of-Speech Tagger (Toutanova et al., 2003).

In the end, the strategies most difficult to identify are *knowledge inference* and *bridging*, for which semantic similarities have to be computed. An inferred concept is a non-paraphrased word for which the following three semantic distances were computed: the distance from word w_1 from the

verbalization to the closest word w_2 from the initial text (expressed in terms of semantic distances in ontologies, LSA and LDA) and the distances from both w_1 and w_2 to the textual fragments in-between consecutive self-explanations. The latter distances had to be taken into consideration for better weighting the importance of each concept, with respect to the whole text. In the end, for classifying a word as inferred or not, a weighted sum of the previous three distances is computed and compared to a minimum imposed threshold which was experimentally set at 0.4 for maximizing the precision of the knowledge inference mechanism on the used sample of verbalizations.

As bridging consists of creating connections between different textual segments from the initial text, cohesion was measured between the verbalization and each sentence from the referenced reading material. If more than 2 similarity measures were above the mean value and exceeded a minimum threshold experimentally set at 0.3, bridging was estimated as the number of links between contiguous zones of cohesive sentences. Compared to the knowledge inference threshold, the value had to be lowered, as a verbalization had to be linked to multiple sentences, not necessarily cohesive one with another, in order to be considered bridging. Moreover, the consideration of contiguous zones was an adaptation with regards to the manual annotation that considered two or more adjacent sentences, each cohesive with the verbalization, members of a single bridged entity.

We ran an experiment with pupils aged from 9 to 11 who had to read aloud a 450 word-long story, Matilda by Dahl (2007), and to stop in-between at six predefined markers and explain what they understood up to that moment. Their explanations were first recorded and transcribed, then annotated by two human experts (PhD in linguistics and in psychology), and categorized according to scoring scheme. Disagreements were solved by discussion after evaluating each self-explanation individually. In addition, automatic cleaning had to be performed in order to process the phonetic-like transcribed verbalizations.

Verbalizations from 12 pupils were transcribed and manually assessed as a preliminary validation. The results for the 72 verbalization extracts in terms of precision, recall and F1-score are as follows: *causality* ($P = .57$, $R = .98$, $F = .72$), *control* ($P = 1$, $R = .71$, $F = .83$), *paraphrase* ($P = .79$, $R = .92$, $F = .85$), *inferred knowledge* ($P = .34$, $R = .43$, $F = .38$) and *bridging* ($P = .45$, $R = .58$, $F = .5$). As expected, paraphrases, control and causality occurrences were much easier to identify than information coming from pupils' experience (Graesser et al., 1994).

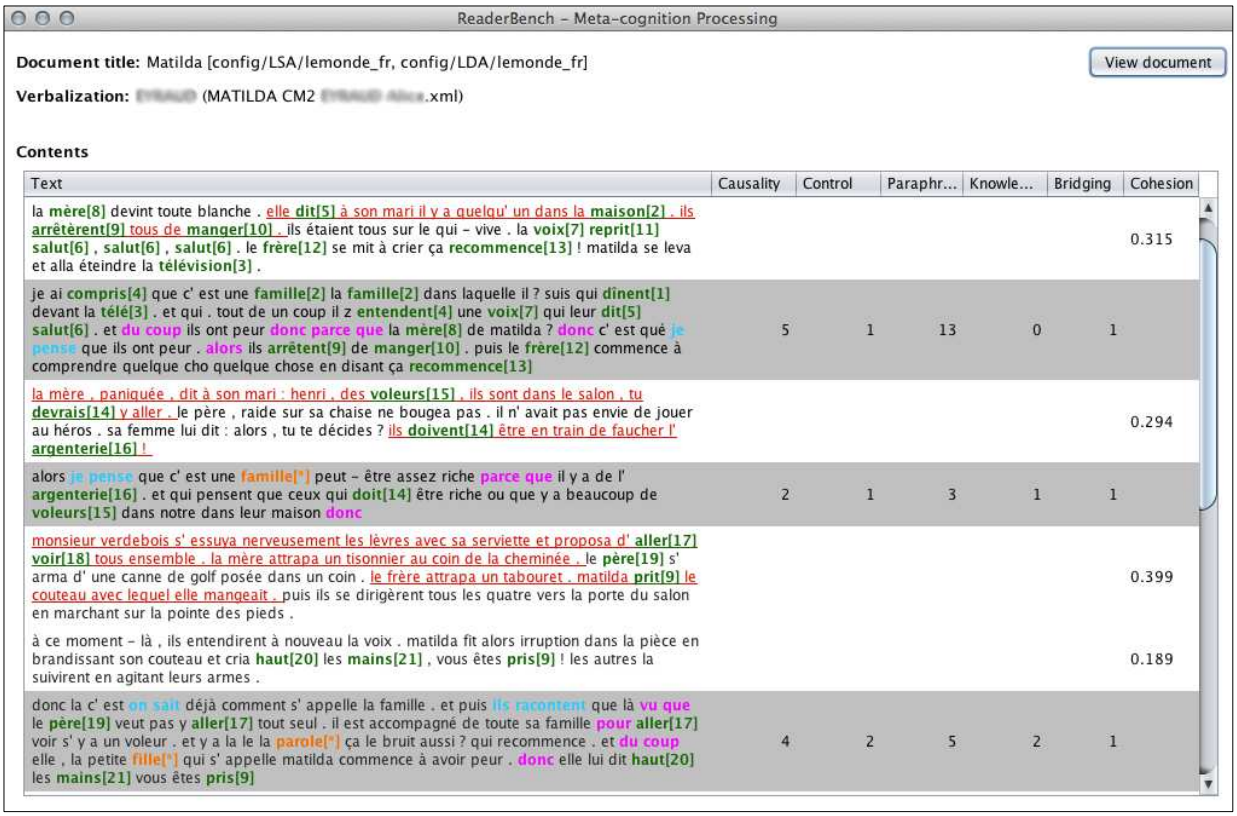


Figure 54. *ReaderBench* (2) Visualization of automatically identified reading strategies. The grey sections represent the pupil's self-explanations, whereas the white blocks represent paragraphs from "Matilda" by Dahl (2007). Causality, control and inferred concepts (that through their definition are not present within the original text) are highlighted only in the verbalization, whereas paraphrases are coded in both the self-explanation and the initial text for a clear traceability of lexical proximity or identity. Bridging, if present, is highlighted only in the original text for pinpointing out the textual fragments linked together through cohesion in the pupil's meta-cognition

Figure 54 depicts the cohesion measures with previous paragraphs from the story in the last column and the identified reading strategies for each verbalization marked in the grey areas, coded as follows: control, causality, paraphrasing [index referred word from the initial text], inferred concept [*] and bridging over the inter-linked cohesive sentences from the reading material. The initial text of the verbalization, including the corresponding manual coding scheme, can be found in Appendix D – Input Examples, Sample Verbalization.

Moreover we have identified multiple particular cases in which both approaches (human and automatic) covered a partial truth that in the end is subjective to the evaluator. For instance, many causal structures close to each other, but not adjacent, were manually coded as one, whereas the system considers each of them separately. For example, "*fille*" ("daughter") does not appear in the text and is directly linked to the main character, therefore marked as an inferred concept by *ReaderBench*, while the evaluator considered it as a synonym. Additionally, when looking at manual

assessments, discrepancies between evaluators were identified due to different understandings and perceptions of pupil's intentions expressed within their metacognitions. Nevertheless, our aim was to support tutors and the results are encouraging (correlated also with the previous precision measurements and with the fact that a lot of noise existed in the transcriptions), emphasizing the benefits of a regularized and deterministic process of identification.

As extensions, we are envisioning two directions: 1/ *generalizing the evaluations* to the whole corpus of pupils' metacognitions (84 verbalizations), but this is a time-consuming process as manual adjustments need to be made to the transcribed verbalizations (e.g., adding punctuation signs in order to facilitate parsing) and 2/ *building an automatic classification model* based on Support Vector Machines (Cortes & Vapnik, 1995) in order to *predict* the *comprehension level* of each learner based on his/her reading strategies; post-tests were administered to each pupil, comprehension scores were manually determined using these tests/questionnaires and our aim is to estimate a comprehension level class using as inputs the automatically identified reading strategies.

8.2 Textual Complexity Analysis Model

Assessing textual complexity can be considered a difficult task due to different reader perceptions primarily caused by prior knowledge and experience, cognitive capability, motivation, interests or language familiarity (for non-native speakers) (see *2.1.3* Cohesion and Coherence versus Textual Complexity and *2.2* Textual Complexity). Nevertheless, from the tutor perspective, the task of identifying accessible materials plays a crucial role in the learning process since inappropriate texts, either too simple or too difficult, can cause learners to quickly lose interest.

In this context, we propose a multi-dimensional analysis of textual complexity, covering a multitude of factors integrating classic readability formulas, surface metrics derived from automatic essay grading techniques, morphology and syntax factors (Dascalu et al., 2012), as well as new dimensions focused on semantics (Dascalu, Dessus, et al., in press). In the end, subsets of specific factors are aggregated through the use of Support Vector Machines (Cortes & Vapnik, 1995), which has proven to be the most efficient method (Petersen & Ostendorf, 2009; T. François & Miltsakaki, 2012). In order to provide an overview, the textual complexity dimension, with their corresponding performance scores, are presented in Table 27, whereas the following subsections describe each dimension with its complexity factors.

8.2.1 Surface Analysis

Surface analysis addresses lexical and syntactic levels and consists of measures computed to determine factors like fluency, complexity, readability taking into account lexical and syntactic elements (e.g., words, commas, phrase length, periods).

A Readability

Traditional readability formulas (Brown, 1998) are simple methods for evaluating a text's reading ease based on simple statistical factors as sentence length or word length. Although criticized by discourse analysts (Davison & Kantor, 1982) as being weak indicators of comprehensibility and for not closely aligning with the cognitive processes involved in text comprehension, their simple mechanical evaluation makes them appealing for integration in our model. Moreover, by considering the fact that reading speed, retention and reading persistence are greatly influenced by the complexity of terms and overall reading volume, readability formulas can provide a viable approximation of the complexity of a given text, considering that prior knowledge, personal skills and traits (e.g., intelligence), interest and motivation are at an adequate level or of a similar level for all individuals of the target audience. In addition, the domain of texts, itself, must be similar because subjectivity increases dramatically when addressing cross-domain evaluation of textual complexity.

Starting from simple lexical indicators, numerous mathematical formulas were developed to tackle the issue of readability. The following three measures can be considered the most famous:

- The *Flesch Reading Ease Readability Formula* (see Equation 28) is one of the oldest and most accurate readability formulas, providing a simple approach to assess the grade-level of chat participants or the difficulty of a reading material; the higher the score, the easier the text is considered in terms of reading, not necessarily understanding (Flesch, 1948).

$$RE = 206{,}835 - (1{,}015 * ASL) - (84{,}6 * ASW) \quad (28)$$

Where: *RE* = Readability Ease; *ASL* = Average Sentence Length (the number of words divided by the number of sentences); *ASW* = Average number of Syllables per Word (he number of syllables divided by the number of words).

- The *Gunning's Fog Index* (or FOG) Readability Formula (see Equation 29) is based on the opinion of Gunning (1952) that certain documents were full of "fog" and unnecessary complexity; the index estimates the number of years of education needed to understand the text while reading it for the first time. Although approximating hard words as words with more than two syllables can be seen as a drawback, we chose this estimation due to its simplicity (Gunning, 1952).

$$FOG = (ASL + PHW) * 0{,}4 \quad (29)$$

Where: *ASL* = Average Sentence Length (the number of words divided by the number of sentences); *PHW* = Percentage of Hard Words (in current implementation words with more than 2 syllables and not containing a dash).

- The *Flesch Grade Level Readability Formula* (see Equation 30) rates documents on U.S. grade school level, therefore simplifying the process of assigning certain materials to a targeted grade of pupils/students. As practical applications, this formula is integrated in Microsoft Word and is used as a standard test by the US Government Department of Defense (Kincaid, Fishburne, Rogers, & Chissom, 1975).

$$FKRA = (0{,}39 * ASL) + (11{,}8 * ASW) - 15{,}59 \quad (30)$$

Where: *FKRA* = Flesch-Kincaid Reading Age; *ASL* = Average Sentence Length (the number of words divided by the number of sentences); *ASW* = Average number of Syllable per Word (the number of syllables divided by the number of words).

B Trins and proxes

Page's initial study was centered on the idea that computers can be used to automatically evaluate and grade student essays using only statistically and easily detectable attributes, as effective as human teachers (E. Page, 1966, 1968; Wresch, 1993). In order to perform a statistical analysis, Page correlated two concepts: *proxes* (computer approximations of interest) with *human trins* (intrinsic variables – human measures used for evaluation) for better quantifying an essay's complexity. A correlation of .71 proved that computer programs could predict grades quite reliably, similar to the inter-human correlation. Starting for Page's metrics of automatically grading essays and taking into consideration Slotnick's method (Slotnick, 1972; Wresch, 1993) of grouping proxes based on their

intrinsic values, the following categories were used within our model for estimating textual complexity (see Table 25).

Table 25. *ReaderBench* (2) Surface analysis factors.

Quality	Proxes
Fluency	Normalized number of commas
	Normalized number of words
	Average number of words per sentence
Diction	Average word length
	Average number of syllables per word
	Percent of hard words (extracted from FOG Formula)
Structure	Normalized number of blocks (paragraphs)
	Average block (paragraph) size
	Normalized number of sentences
	Average sentence length

Normalization is inspired from data-mining and information retrieval (Manning et al., 2008) and our results improved by applying the logarithmic function on some of the previous factors in order to smooth results, while comparing documents of different size. All the above proxes determine the average consistency of sentences and adequately model their complexity at surface/lexical level.

C Entropy

Entropy, derived from Information Theory (Shannon, 1948, 1951), models the text in an ergodic manner and provides relevant insight regarding textual complexity at character and word level by ensuring diversity among the elements of the analysis (see Equation 31). The assumption of induced complexity pursues the following hypothesis: a more complex text contains more information and requires more memory and more time for the reader to process. Therefore, disorder modeled through entropy is reflected in the diversity of characters and of word stems used, within our implemented model, as analysis elements. The use of stems instead of actual concepts is argued by their better expression of the root form of related concepts, more relevant when addressing syntactic diversity.

$$H(X) = -\sum_{\substack{c=stemmed\ word \\ or \\ c=character}} p(c)ln(p(c)) \qquad (31)$$

8.2.2 Metrics for word complexity

From a different perspective, word complexity was treated as a combination of the following factors: syllable count, distance between the inflected form, lemma and stem, whereas specificity is reflected in inverse document frequency from the training corpora, the distance in hypernym tree and the word polysemy count from the ontology. As an overview of the entire discourse, all these metrics are computed in a simple manner, by summing up the relevant values for all the words within text (only dictionary words after the initial NLP pipe processing) and then dividing the sum by the total number of words.

The relevance of using the *mean syllable count per word* resides in the intuition that the number of syllables of a word correlates directly with its difficulty. In general, the more syllables a word has, the harder it is to pronounce. When learning a language, for instance, speakers tend to use words with fewer syllables that are easier to say out loud. As the learner's proficiency in a language increases, the usage of more difficult, multisyllabic words also increases. Anyway, although pronunciation is linked to textual complexity, it differs greatly from comprehension in the sense that only a shallow analysis cannot be sufficient to grasp text difficulty (Benjamin, 2012).

In terms of the *mean polysemy count per word*, we operate under the assumption that the more possible senses a word has, the more difficult it would be to use in a text and to correctly identify its sense. Therefore, simpler texts will contain words that are less ambiguous, while more complex texts, on the whole, will use more words with a higher sense count.

The *distance* within the *hypernym tree to the ontology root* can be seen as a measure of word specialization and specificity. In other words, the more elaborated the path to the root of the ontology hierarchy, the more specific the text can be considered, covering more peculiar terms. The farther a word is from the hypernym tree root, the more specialized it is. From a computational perspective, due to multiple possible paths and word senses, we determine this distance using a backtracking algorithm (Cormen et al., 2009).

While addressing the *differences* between the *inflected form,* the *lemma* and the *stem* of a word, it becomes clear that a correlation exists between the complexity of a word's derivation and its overall complexity – as multiple prefixes and suffixes are juxtaposed, the more complex the word can be considered.

8.2.3 Morphology and Syntax

A Complexity, Accuracy and Fluency

Complexity, accuracy, and fluency (CAF) measures of texts have been used in linguistic development and in second language acquisition (SLA) research (House & Kuiken, 2009). *Complexity* captures the characteristic of a learner's language, reflected in a wider range of vocabulary and grammatical constructions, as well as communicative functions and genres (Schulze, 2010). *Accuracy* highlights a text's conformation to our experience with other texts, while *fluency*, in oral communication, captures the actual volume of text produced in a certain amount of time. Similar to the previous factors, these measures play an important role in automated essay scoring and textual complexity analysis. Schulze (2010) considered that selected *complexity* measures should be divided into two main facets of textual complexity: sophistication (richness) and diversity (variability of forms). The defined measures depend on six units of analysis: letter (*l*), word form (*w*), bigram (*b* – groups of two words) and period unit (*p*), word form types (*t*) and unique bigrams (*u*). Additionally, textual complexity is devised into lexical and syntactic complexity:

Lexical Complexity:

- *Diversity* is measured using Carroll's Adjusted Token Type Ratio (see Equation 32) (Schulze, 2010).

$$v_1 = \frac{t}{\sqrt{2w}}, with\ \frac{1}{\sqrt{2w}} \leq v_1 \leq \sqrt{\frac{w}{2}} \quad (32)$$

- *Sophistication* estimates the complexity of a word's form in terms of average number of characters (see Equation 33) (Schulze, 2010).

$$v_2 = \frac{l}{w}, with\ 1 \leq v_2 \leq l \quad (33)$$

Syntactic Complexity:

- *Diversity* captures syntactic variety at the smallest possible unit of two consecutive word forms (see Equation 34). Therefore Token Type Ratio is also used, but at a bigram level (Schulze, 2010).

$$v_3 = \frac{u}{\sqrt{2b}}, with\ \frac{1}{\sqrt{2b}} \leq v_3 \leq \sqrt{\frac{b}{2}} \quad (34)$$

- *Sophistication* is expressed in terms of mean number of words per period unit length and its intuitive justification is that longer clauses are, in general, more complex than short ones (see Equation 35) (Schulze, 2010).

$$FKRA = (0{,}39 * ASL) + (11{,}8 * ASW) - 15{,}59 \quad (35)$$

All the previous measures can be integrated into a unique measure of textual complexity at lexical and syntactic levels. Following this idea, these factors were balanced by computing a rectilinear distance (Raw Complexity, RC) as if the learner had to cover the distance along each of these dimensions. Therefore, in order to reach a higher level of textual complexity, the learner needs to improve on all four dimensions (see Equation 36) (Schulze, 2010).

$$RC = \left|v_1 - \frac{1}{\sqrt{2w}}\right| + |v_2 - 1| + \left|v_3 - \frac{1}{\sqrt{2b}}\right| + |v_4 - 1| \quad (36)$$

Afterwards, CAF is computed as a balanced complexity by subtracting the range of the four complexity measures (max – min) from the raw complexity measure (see Equation 37).

$$CAF = RC - (max(v_1, v_2, v_3, v_4) - min(v_1, v_2, v_3, v_4)) \quad (37)$$

The ground argument for this adjustment is that if one measure increases too much, it will always be to the detriment of another. Therefore, the measure of raw complexity is decreased by a large amount if the four vector measures vary widely and by a small amount if they are very similar. Moreover, the defined measure captures lexical and syntactic complexity evenly, provides two measures for sophistication and two measures for diversity and, in the end, compensates for large variations of the four vector measures.

B Part-of-Speech Statistics and Parsing Tree Structure

Starting from different linguistic categories of lexical items, our aim is to convert morphological information regarding the words and the sentence structure into relevant metrics to be assessed in order to better comprehend textual complexity. In this context, parsing and part of speech (POS) tagging play an important role in the morphological analysis of texts, in terms of textual complexity,

by providing two possible vectors of evaluation: the normalized frequency of each part of speech and the structural factors derived from the parsing tree. Although the most common parts of speech used in discourse analysis are nouns and verbs, our focus was aimed at prepositions, adjectives and adverbs that dictate a more elaborate and complex structure of the text. Moreover, pronouns, that through their use indicate the presence of co-references, also indicate a more intertwined and complex structure of the discourse. On the other hand, multiple factors can be derived from analyzing the structure of the parsing tree: an increased number of leafs, a greater overall size of the tree and a higher maximum depth indicate a more complex structure, therefore an increased textual complexity (Gervasi & Ambriola, 2002).

8.2.4 Semantics

Firstly, as seen in *2.1* Coherence and Comprehension, *textual complexity* is linked to *cohesion* in terms of comprehension; in other words, in order to understand a text, the reader must first create a well-connected representation of the information withheld, a situation model (van Dijk & Kintsch, 1983) (see Figure 2). This connected representation is based on linking related pieces of textual information that occur throughout the text. Therefore, cohesion reflected in the strength of inner-block and inter-block links extracted from the cohesion graph influences readability, as semantic similarities govern the understanding of a text. In this context, discourse cohesion is evaluated at a macroscopic level as the average value of all links in the constructed cohesion graph (Trausan-Matu, Dascalu, & Dessus, 2012; Dascalu, Dessus, et al., in press).

Secondly, a variety of metrics based on the *span* and the *coverage of lexical chains* (Galley & McKeown, 2003) provide insight in terms of lexicon variety and of cohesion, expressed in this context as the semantic distance between different chains. Moreover, we imposed a threshold of minimum of 5 words per lexical chain in order to consider it relevant in terms of overall discourse; this value was determined experimentally after running simulations with increasing values and observing the correlation with predefined textual complexity levels.

Thirdly, *entity-density features* proved to influence readability as the number of entities introduced within a text is correlated to the working memory of the text's targeted readers. In general, entities consisting of general nouns and named entities (e.g., people's names, locations, organizations) introduce conceptual information by identifying, in most cases, the background or the context of

the text. More specifically, entities are defined as a union of named entities and general nouns (nouns and proper nouns) contained in a text, with overlapping general nouns removed. These entities have an important role in text comprehension due to the fact that established entities form basic components of concepts and propositions on which higher level discourse processing is based (Feng, Jansche, Huenerfauth, & Elhadad, 2010). Therefore, the entity-density factors focus on the following statistics: the number of entities (unique or not) per document or sentence, the percentages of named entities per document, the percentage of overlapping nouns removed or the percentage of remaining nouns in total entities.

Finally, another dimension focuses on the ability to resolve *referential relations* correctly (Raghunathan et al., 2010; Lee et al., 2011; Lee et al., 2013) as *co-reference inference* features also impact comprehension difficulty (e.g., the overall number of chains, the inference distance or the span between concepts in a text, number of active co-reference chains per word or per entity).

8.2.5 Combining Textual Complexity Factors through Support Vector Machines

All the measures previously defined capture in some degree different properties of the analyzed text (readability, fluency, language diversity and sophistication, morphological structure, cohesion, etc.) and therefore can be viewed as attributes that describe the text. In order to use these attributes to estimate the complexity of the text, we have used a classifier that accepts as inputs text attributes and outputs the minimum grade level required by a reader to comprehend the specified text. In our integrated textual complexity analysis model we have opted for Support Vector Machine (SVM) classifiers that have been proven to be the most appropriate (Petersen & Ostendorf, 2009; T. François & Miltsakaki, 2012). A SVM (Cortes & Vapnik, 1995; Press, Teukolsky, Vetterling, & Flannery, 2007) is typically a binary linear classifier that maps the input texts seen as *d*-dimensional vectors to a higher dimensional space (hyperspace) through the mapping of a kernel function, in which, hopefully, these vectors are linearly separable by a hyperplane (see Figure 55).

Due to the fact that binary classifiers can map objects only into two disjoint classes, our multiclass problem can be solved using multiple Support Vector Machines, each classifying a category of texts with different predefined classes of complexity (C.-W. Hsu & Lin, 2002; Duan & Keerthi, 2005). A one-versus-all approach implementing the winner-takes-all strategy is used to deal with the

problem of multiple SVM returning 1 for a specific text (the classifier with the highest output function assigns the class).

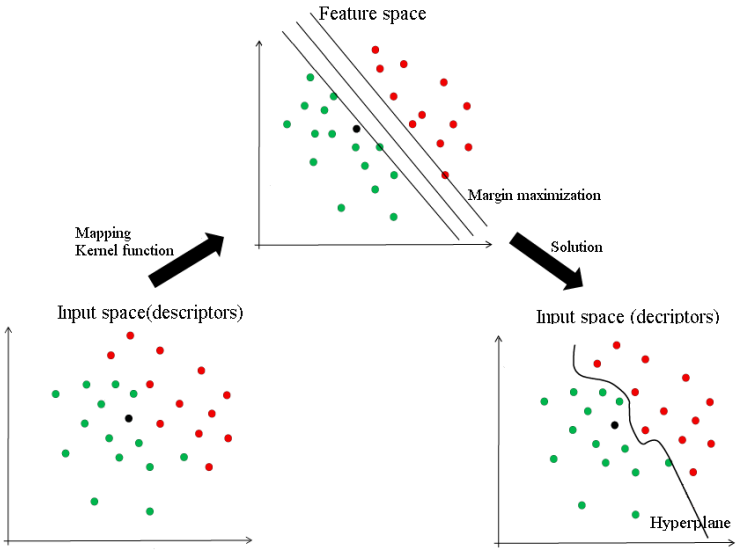


Figure 55. General binary SVM mapping and separation through a hyperplane – adapted from Kozak, Agrawal, Machuy, and Csucs (2009).

LIBSVM (C.-C. Chang & Lin, 2011) was used to ease the implementation of the classifier and integrated in *ReaderBench*. An RBF kernel with degree 3 was selected and a Grid Search method (C. W. Hsu, Chang, & Lin, 2010; Bergstra & Bengio, 2012) was enforced to increase the effectiveness of the SVM through the parameter selection process for the Gaussian kernel. Exponentially growing sequences for C and γ were used ($C \in \langle 2^{-5}, 2^{-3}, \dots, 2^{13}, 2^{15} \rangle$, $\gamma \in \langle 2^{-15}, 2^{-13}, \dots, 2^{1}, 2^{3} \rangle$) and each combination of parameter choices was checked using the testing corpora; in the end, the parameters with the best precision were selected.

8.2.6 Validation of the Integrated Textual Complexity Analysis Model

In order to train our complexity model, we have opted to automatically extract English texts from TASA, using its Degree of Reading Power (DRP) score, into six classes of complexity (McNamara et al., in press) of equal frequency, as no corpus was available for French (see Table 26).

Table 26. Ranges of the DRP scores as a function of defining the six textual complexity classes (after McNamara et al., in press).

Complexity Class	Grade Range	DRP Minimum	DRP Maximum
1	K-1	35.38	45.99
2	2-3	46.02	51.00
3	4-5	51.00	56.00
4	6-8	56.00	61.00
5	9-10	61.00	64.00
6	11-CCR	64.00	85.80

This validation scenario consisting of approximately 1,000 documents was twofold: we wanted, on one hand, to prove that the complete model is adequate and reliable and, on the other, to demonstrate that high level semantic features provide relevant insight that can be used for automatic classification. In the end, *k*-fold cross validation (Geisser, 1993) was applied for extracting the following performance features (see Table 27 and Figure 56): precision or exact agreement (EA) and adjacent agreement (AA) (T. François & Miltsakaki, 2012), as the percent to which the SVM was close to predicting the correct classification.

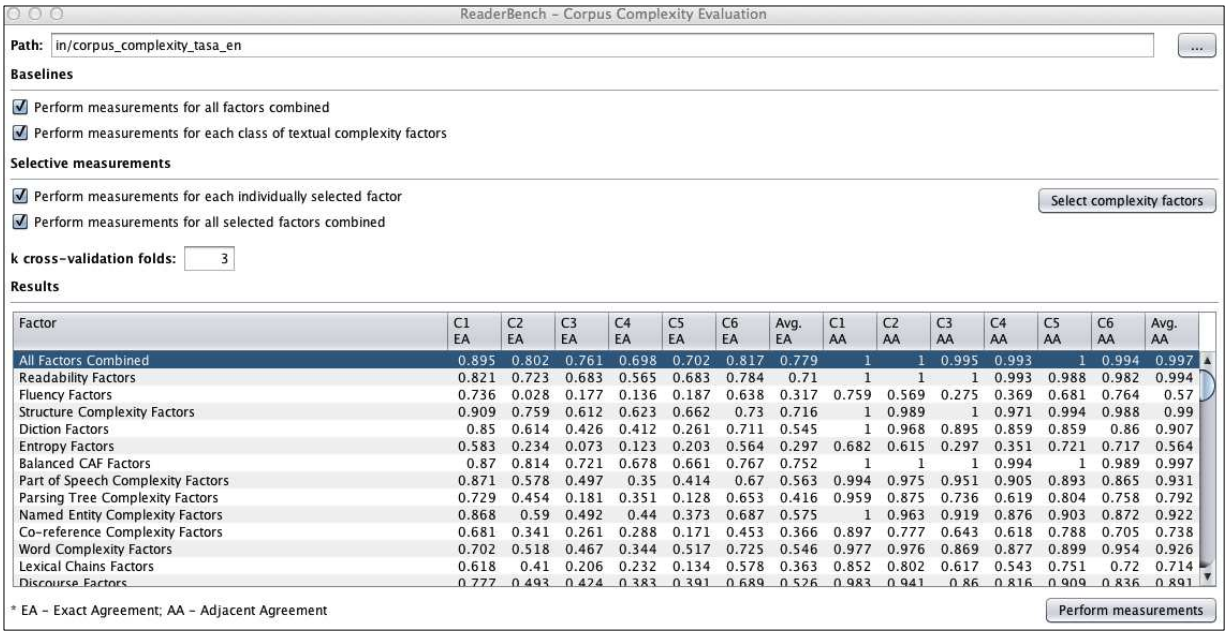
Table 27. *ReaderBench* (2) Textual complexity dimensions.

Depth of metrics	Factors for evaluation	Avg. EA	Avg. AA
Surface Analysis	Readability formulas	.71	.994
	Fluency factors	.317	.57
	Structure complexity factors	.716	.99
	Diction factors	.545	.907
	Entropy factors (words vs. characters)	.297	.564
	Word complexity factors	.546	.926
Morphology & Syntax	Balanced CAF (Complexity, Accuracy, Fluency)	.752	.997
	Specific POS complexity factors	.563	.931
	Parsing tree complexity factors	.416	.792
Semantics	Cohesion through lexical chains, LSA and LDA	.526	.891
	Named entity complexity factors	.575	.922
	Co-reference complexity factors	.366	.738
	Lexical chains	.363	.714

By considering the granular factors, although simple in nature, readability formulas, the average number of words per sentence, the average length of sentences/words and balanced CAF provided the best alternatives at lexical and syntactic level; this was expected as the DRP score is based solely on shallow evaluation factors. From the perspective of word complexity factors, the average polysemy count and the average word syllable count correlated well with the DRP scores. In terms of parts of speech tagging, nouns, prepositions and adjectives had the highest correlation of all types of parts of speech, whereas depth and size of the parsing tree provided also a good insight of textual complexity.

In contrast, semantic factors taken individually had lower scores because the evaluation process at this level is mostly based on cohesive or semantic links between analysis elements and the variance between complexity classes is lower in these cases. Moreover, while considering the evolution from the first class of complexity to the latest, these semantic features don't necessarily have an upward gradient; this can fundamentally affect a precise prediction if the factor is taken into consideration individually. Only 2 entity-density factors had better results, but their values are directly connected to the underlying part of speech (noun) that had the best EA and AA of all morphology factors. Also, the most difficult classes to identify were the second and the third because the differences between them were less noteworthy. The complete results list for all evaluation factors, with detailed information for each dimension, is presented in Appendix C – Textual Complexity.

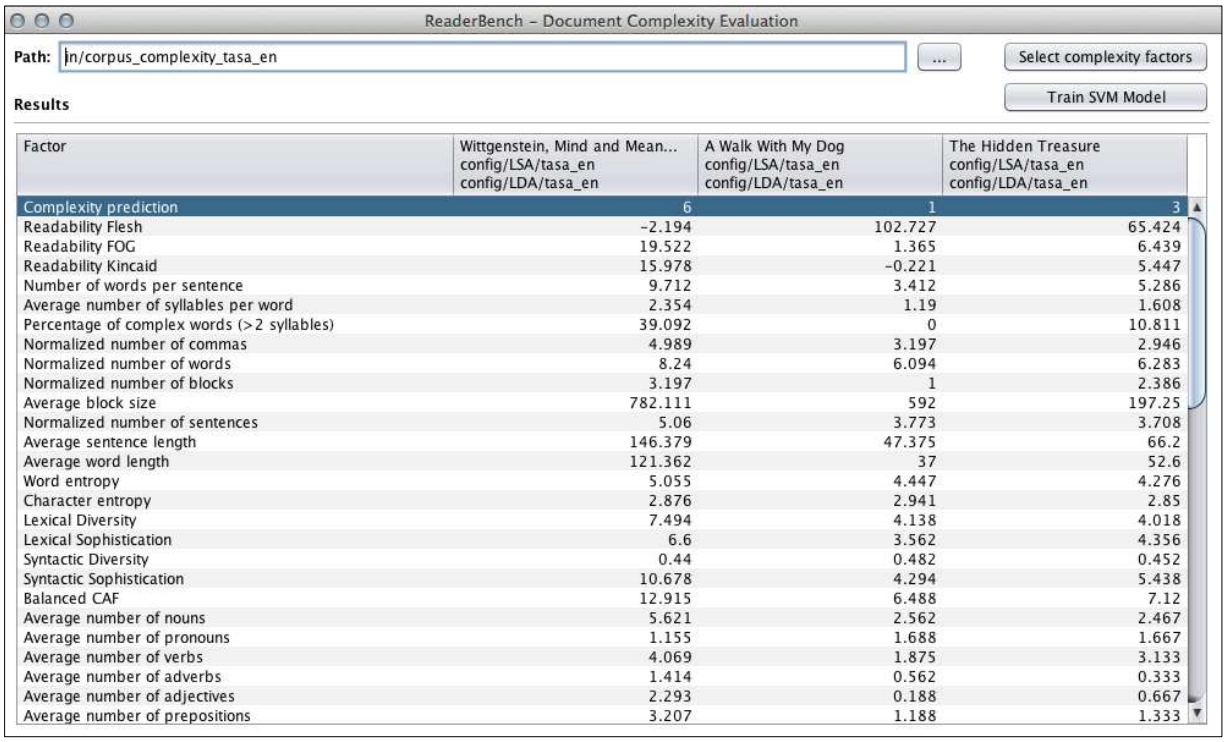
Moreover, besides the factors presented in detail in Dascalu et al. (2012) that were focused on a more shallow approach, of particular interest was how semantic factors correlate to classic readability measures (Dascalu, Dessus, et al., in press). In this context, two additional measurements were performed. Firstly, an integration of all metrics from all textual complexity dimensions proved that the SVMs results are compatible with the DRP scores (EA = .779 and AA = .997), and that they provide significant improvements as they outperform any individual dimension precisions. The second measurement (EA = .597 and AA = .943) used only morphology and semantic measures in order to avoid a circular comparison between factors of similar complexity, as the DRP score is based on shallow factors. This result showed a link between low-level factors (also used in the DRP score) and in-depth analysis factors, which can also be used to accurately predict the complexity of a reading material.



Factor	C1 EA	C2 EA	C3 EA	C4 EA	C5 EA	C6 EA	Avg. EA	C1 AA	C2 AA	C3 AA	C4 AA	C5 AA	C6 AA	Avg. AA
All Factors Combined	0.895	0.802	0.761	0.698	0.702	0.817	0.779	1	1	0.995	0.993	1	0.994	0.997
Readability Factors	0.821	0.723	0.683	0.565	0.683	0.784	0.71	1	1	1	0.993	0.988	0.982	0.994
Fluency Factors	0.736	0.028	0.177	0.136	0.187	0.638	0.317	0.759	0.569	0.275	0.369	0.681	0.764	0.57
Structure Complexity Factors	0.909	0.759	0.612	0.623	0.662	0.73	0.716	1	0.989	1	0.971	0.994	0.988	0.99
Diction Factors	0.85	0.614	0.426	0.412	0.261	0.711	0.545	1	0.968	0.895	0.859	0.859	0.86	0.907
Entropy Factors	0.583	0.234	0.073	0.123	0.203	0.564	0.297	0.682	0.615	0.297	0.351	0.721	0.717	0.564
Balanced CAF Factors	0.87	0.814	0.721	0.678	0.661	0.767	0.752	1	1	1	0.994	1	0.989	0.997
Part of Speech Complexity Factors	0.871	0.578	0.497	0.35	0.414	0.67	0.563	0.994	0.975	0.951	0.905	0.893	0.865	0.931
Parsing Tree Complexity Factors	0.729	0.454	0.181	0.351	0.128	0.653	0.416	0.959	0.875	0.736	0.619	0.804	0.758	0.792
Named Entity Complexity Factors	0.868	0.59	0.492	0.44	0.373	0.687	0.575	1	0.963	0.919	0.876	0.903	0.872	0.922
Co-reference Complexity Factors	0.681	0.341	0.261	0.288	0.171	0.453	0.366	0.897	0.777	0.643	0.618	0.788	0.705	0.738
Word Complexity Factors	0.702	0.518	0.467	0.344	0.517	0.725	0.546	0.977	0.976	0.869	0.877	0.899	0.954	0.926
Lexical Chains Factors	0.618	0.41	0.206	0.232	0.134	0.578	0.363	0.852	0.802	0.617	0.543	0.751	0.72	0.714
Discourse Factors	0.777	0.493	0.424	0.383	0.391	0.689	0.526	0.983	0.941	0.86	0.816	0.909	0.836	0.891

Figure 56. *ReaderBench* (2) Textual complexity evaluation.
Starting from a pre-processed corpus, the user has the opportunity to perform the following measurements applied on: 1/ the complete SVM model with all factors integrated; 2/ each individual complexity dimension (a predefined subset of textual complexity metrics); 3/ a specific set of selected complexity factors, on which individual measurements or a single combined evaluation can be performed. In the end, a table is automatically generated including the used factor (individual, textual complexity dimension or specific aggregation), exact and adjacent agreements for each complexity class from the corpus, as well as the average agreement values

In terms of usability, besides the possibility to train and evaluate new textual complexity models on a given corpora (see Figure 56), *ReaderBench* enables tutors to assess the complexity of new reading materials based on the selected complexity factors and a pre-assessed corpus of texts, pertaining to different complexity dimensions. By comparing multiple loaded documents, tutors can better grasp each evaluation factor, refine the model to best suit their interests in terms of the targeted measurements and perform new predictions using only their features (see Figure 57).



Factor	Wittgenstein, Mind and Mean... config/LSA/tasa_en config/LDA/tasa_en	A Walk With My Dog config/LSA/tasa_en config/LDA/tasa_en	The Hidden Treasure config/LSA/tasa_en config/LDA/tasa_en
Complexity prediction	6	1	3
Readability Flesh	-2.194	102.727	65.424
Readability FOG	19.522	1.365	6.439
Readability Kincaid	15.978	-0.221	5.447
Number of words per sentence	9.712	3.412	5.286
Average number of syllables per word	2.354	1.19	1.608
Percentage of complex words (>2 syllables)	39.092	0	10.811
Normalized number of commas	4.989	3.197	2.946
Normalized number of words	8.24	6.094	6.283
Normalized number of blocks	3.197	1	2.386
Average block size	782.111	592	197.25
Normalized number of sentences	5.06	3.773	3.708
Average sentence length	146.379	47.375	66.2
Average word length	121.362	37	52.6
Word entropy	5.055	4.447	4.276
Character entropy	2.876	2.941	2.85
Lexical Diversity	7.494	4.138	4.018
Lexical Sophistication	6.6	3.562	4.356
Syntactic Diversity	0.44	0.482	0.452
Syntactic Sophistication	10.678	4.294	5.438
Balanced CAF	12.915	6.488	7.12
Average number of nouns	5.621	2.562	2.467
Average number of pronouns	1.155	1.688	1.667
Average number of verbs	4.069	1.875	3.133
Average number of adverbs	1.414	0.562	0.333
Average number of adjectives	2.293	0.188	0.667
Average number of prepositions	3.207	1.188	1.333

Figure 57. *ReaderBench* (2) Document complexity evaluation.

Based on a pre-trained corpus, the user selects the complexity factors to be automatically used within the SVM model (by default all factors are pre-selected) and *ReaderBench* generates a complexity prediction for each loaded document, as well as all values corresponding to the selected individual factors in order to have a comparison of the evolution of specific metrics between different documents

8.3 Comparison of *ReaderBench* to *iSTART, Dmesure and Coh-Metrix*

This section addresses in extent the comparison between *ReaderBench* and 3 systems that seemed most close to its goals: *iStart* in terms of reading strategies (see *2.3* Reading Strategies), whereas *Dmesure* and *Coh-Metrix* are representative for textual complexity (see *2.2* Textual Complexity).

Table 28. *ReaderBench* versus *iSTART* (O'Reilly et al., 2004; Graesser et al., 2005; McNamara, Boonthum, et al., 2007).

Benefits of *ReaderBench*	Benefits of *iStart*
Educational perspective	
Adaptation of the proposed methodology to the specificity of the undergone experiments	Initial methodology designed for assessing reading comprehension
Refinement of the reading strategies in terms of the observed pupil's behavior (no prediction, elaboration was generalized to knowledge inference)	Initial taxonomy of reading strategies
Separate identification of reading strategies and a more fine-grained comparison to the gold standard, without a direct liaison to predicting learner comprehension	Assignment of an overall relevance score on a [1; 4] scale, easily linkable to comprehension
The evaluation targeted primary school pupils – elliptical expressions, pauses and repetitions in oral speech that impacted the transcription process	Analysis of student self-explanations – adequate and coherent language, direct recording of textual representation
Retrospective view, with focus on accurate identification of different strategies	Proactive perspective, with emphasis on the impact of the system on students' comprehension
Tutor inquiry oriented analysis, with accent on the demarcation of different strategies	The use of different animated agents to present a warmer, more interactive and more user friendly perspective of the analysis
Technical perspective	
In-depth methods of extracting reading strategies using multiple heuristics (word- and LSA- heuristics were analyzed in the first two studies, later refined in *ReaderBench*)	Word-based and LSA centered extraction of strategies
French corpus, much more difficult to analyze in terms of natural language processing; moreover, the system enables applying the NLP pipe to both French and English texts	English self-explanations analyzed within a web-form, with no NLP specific processing
Preprocessing and cleaning of verbalizations was required after manual phonetic transcription	

Table 29. *ReaderBench* versus *Dmesure* (T. François, 2012; T. François & Miltsakaki, 2012).

Benefits of *ReaderBench*	Benefits of *Dmesure*
Educational perspective	
Broad view covering multiple analysis levels, from surface analysis to semantics	Focalized analysis, granting a comprehensive view of lexical, syntactic and morphological factors
Shift of perspective towards demonstrating that high – level factors can be also used to accurately predict the complexity of a document	
Technical perspective	
Integration of a complete NLP pipe for both French and English	Application of specific NLP techniques, but limited due to the use *TreeTagger* (H. Schmidt, 1995), a language independent parser
Integration of the most commonly used factors, plus a multitude of new factors extracted from the cohesion graph	Exhaustive analysis of possible factors (more than 300 factors), therefore enhancing the chance of accurately predicting the complexity class by combining multiple inputs; similar to some extent to Kukemelk and Mikk (1993) regarding the spread of statistics; mostly surface, lexical and morphological factors, with only two factors derived from LSA
The use of solely SVMs for classifying documents as multiple studies consider them the most accurate classifiers, efficient also when addressing non-linear separable variables	A comprehensive analysis of multiple classification algorithms
Intuitive user interface, enabling the training and the evaluation of a new textual complexity model based on the factors selected by the user, plus a comparison of different document features	No visual interface
1,000 documents used for training the SVM; Drawback: the comparison was made using the DRP scores from TASA	FFL corpus, manually annotated, which greatly improved the overall relevance of the analysis
Greater agreement values and near perfect adjacent agreement, as results are compared to automatic scores that induced a normalization of the initial documents classification; experiments performed on approx. 250 online reading assignments (Dascalu et al., 2012) proved that correlations dramatically decrease when using inconsistent initial classifications	Lower scores, meaningful nevertheless and completely justifiable while considering the used corpus and its specificity

Table 30. *ReaderBench* versus *Coh-Metrix* (Graesser et al., 2004; McNamara et al., 2010).

Benefits of *ReaderBench*	Benefits of *Coh-Metrix*
Educational perspective	
Explicit extraction of reading strategies and assessment of textual complexity using cohesion as a central measure (ingoing links with regards to cohesion)	Emphasis on coherence from which multiple analysis dimension emerge (outgoing links from coherence)
Extensible cohesion-based model applicable to both general texts and CSCL conversations, more specifically chats and forum discussion threads	
Technical perspective	
Multi-hierarchical analysis, integrating multiple natural language analysis techniques	Extensive use of LSA and of other relevant measures
Internal discourse structure built as the cohesion graph	Most commonly, similarity is expressed as LSA cosine similarity between adjacent analysis elements
Broader view, integrating factors identified as adequate within other studies	A more detailed analysis of possible factors, covering more scenarios
	Aggregation of results and visualization of multiple graphs

9 *ReaderBench* (3) – Involvement and Collaboration Assessment through Cohesion and Dialogism

Although participants' involvement in chat environments has been studied in previous systems, as mentioned in Overview of Empirical Studies, *ReaderBench* has brought a series of remarkable improvements in terms of collaborative learning:

- Emphasis and better support of the dialogical and polyphonic model previously proposed in *PolyCAFe* with new visualizations and evaluation factors.
- Refinement of the initial collaboration assessment model (Dascalu, Rebedea, et al., 2010; Trausan-Matu, Dascalu, & Rebedea, 2012) based on the social knowledge-building effect, through the use of the cohesion graph (Trausan-Matu, Dascalu, & Dessus, 2012; Dascalu, Trausan-Matu, et al., in press).
- A novel collaboration evaluation model based on the overlapping effect of voices seen as semantic chains (see *7.5* Dialogism and Voice Inter-Animation) pertaining to different participants.
- The validation of the evaluation mechanics on a long-term discussion group, seen as an aggregation of multiple threads across a longer timespan, and not only the assessment of individual chat conversations (Nistor et al., 2013; Nistor et al., in press; Nistor, Baltes, et al., submitted; Nistor, Dascalu, et al., submitted).

9.1 Participant Involvement Evaluation

Besides the identification of topics in the discussion for each participant, significant for pinpointing out the covered concepts, *ReaderBench* also supports participant interaction modeling covering a deeper qualitative dimension, obtained by considering the utterance scores (see *7.4* Cohesion-based Scoring Mechanism). Internally, an interaction graph is built with participants as nodes and the

weight of links equal to the sum of interventions scores multiplied by the cohesion function with the referred element of analysis, extracted from the cohesion graph. Therefore, by performing social network analysis (see *3.2* Social Network Analysis) on the previous participant interaction graph, the scale of analysis is shifted towards an individual perspective, centered on each of the participants. In the end, the size of each node in the interaction graph is directly proportional to its corresponding betweenness score (Brandes, 2001; Bastian et al., 2009). Due to the fact that for chat conversations we are dealing in most cases with a complete graph in which the betweenness score for all nodes is 0, participants are displayed as points (see Figure 58). As cohesive links can exist between utterances pertaining to the same speaker, the visualization also includes the inner links equal to the importance of the utterances expressed as a continuation of the discourse, pertaining to the same participant; for some conversations, these values can be comparable in strength to the sum of all other out-going links, marking an individual behavior instead of collaboration. Similar mechanics, when employed on a larger discussion group or community obtained from an aggregation of multiple conversations (chat sessions or forum discussion threads), become more meaningful and provide a clearer global perspective of the interactions between participants (see *9.3* Long-term Discussion Groups Evaluation). Moreover a clear separation must be made: personal involvement is expressed as the cumulative utterance importance scores, whereas the interaction graph reflects the exchange of information through cohesive links, making the two perspectives complementary one to another.

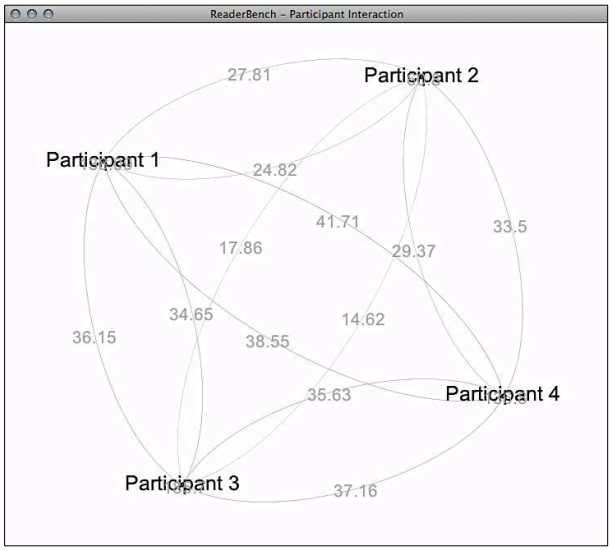


Figure 58. *ReaderBench* (3) Participant centered view of the interaction graph.
The strength of the link between two speakers is reflected in the cumulated effect of each intervention measured through its importance score and reflected in cohesion. In case of a chat conversation with a reduced number of participants, it is most likely to obtain in the end a complete graph, in which the betweenness scores for all nodes are equal to 0, implicitly reducing their diameter to 0

Moreover, an evolution graph of each participant's involvement throughout the conversation, similar to the visualizations provided by *Polyphony* (Trausan-Matu, Rebedea, et al., 2007) and *A.S.A.P.* (Dascalu et al., 2008a) (see *5.1 A.S.A.P.* – Advanced System for Assessing Chat Participants) is also generated (see Figure 59.a), useful for observing interaction patterns. For example, zones with a high slope for one participant are usually in the detriment of the involvement of others and represent areas of the conversation dominated by one participant. On the contrary, comparable growths of multiple participants in a given area induce an equitable involvement and possibly, although not mandatory, collaboration seen as building collaborative knowledge among multiple participants. In the particular case presented in Figure 59.b, all the utterances from the conversation transcript with the identifier from 220 up to 235 pertain solely to Participant 3, from 242 to 261 only two interventions do not belong to Participant 4, whereas Participant 1 completely dominated the discussion between 288 up to 300. Therefore the generated graph is clearly useful for highlighting zones with differential involvement of participants in the conversation.

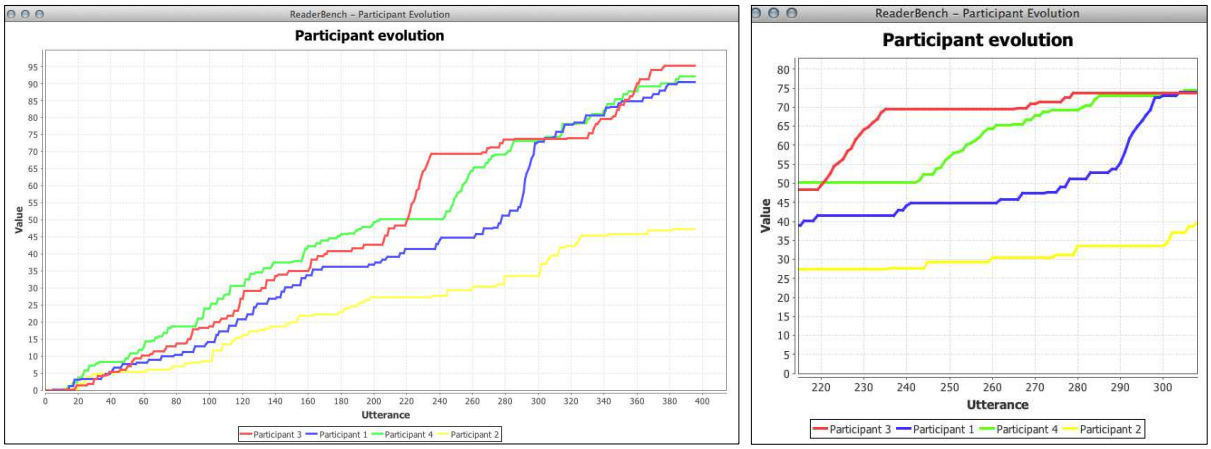


Figure 59. *ReaderBench* (3) Participants' involvement evolution graph.
a. global view of the entire discussion; b. expansion segment of a. around utterance 260

Following the transition from a global view of the discourse to a user-centered perspective, a similar visualization component of the conceptual space for each chat participant as a mind-map, based on semantic similarities between concepts, is generated (see Figure 60 and *7.3* Topics Extraction). Terms central to a given discussion may not appear in any utterance but, nonetheless, be worth displaying for comprehension's sake. We thus enriched the previously identified participant's topics list with inferred concepts, not mentioned within the text, but the actual visualization component has a lowered threshold (30% in this particular case) as more diverse concepts are used throughout the conversation, with a smaller overall cohesion in comparison to reading materials. Moreover, as

the identified list of topics per participant is much more dispersed and has a lower intrinsic cohesion in comparison to a reading material, we opted for eliminating the visualization of the list of inferred concepts as it was misleading; therefore, the inferred concepts are only displayed within the network (see Figure 60).

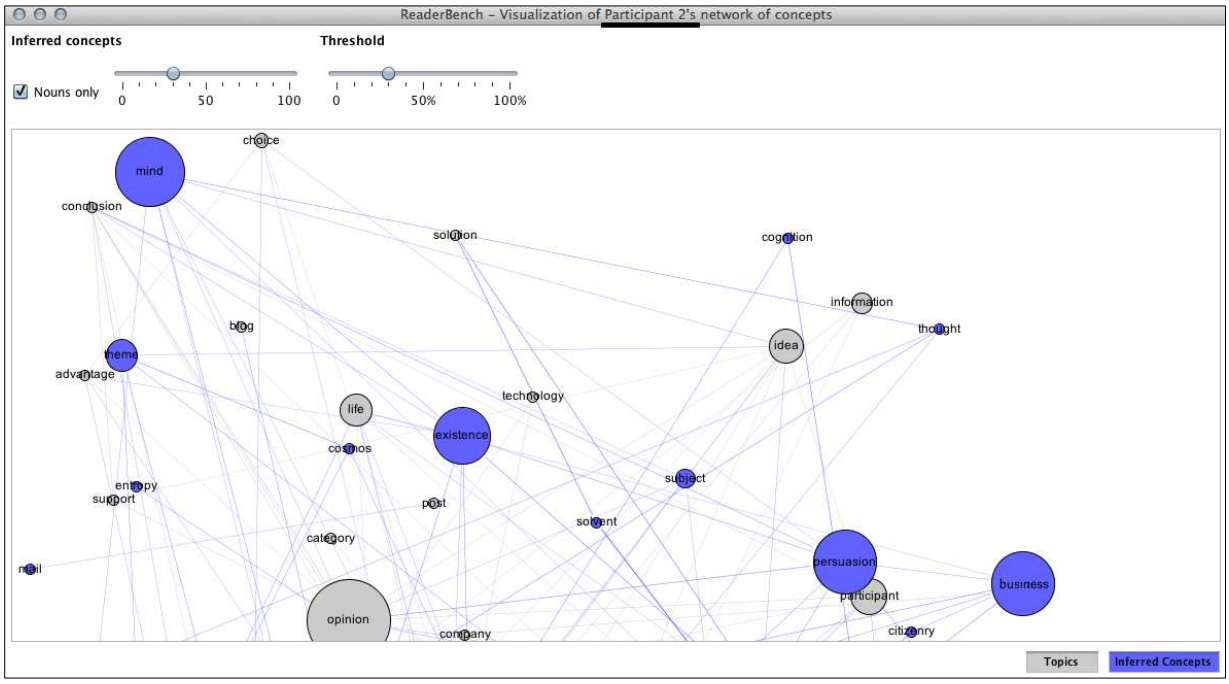


Figure 60. *ReaderBench* (3) Network of concepts generated for a specific participant.

9.2 Collaboration Assessment

In order to thoroughly assess collaboration, we have proposed two computational models. The first model (Dascalu, Trausan-Matu, et al., in press) based on the effect of social knowledge-building, is a refinement of the gain-based collaboration assessment (Dascalu, Rebedea, et al., 2010; Trausan-Matu, Dascalu, & Rebedea, 2012) (see *6.4.2* Collaboration Assessment) and takes full advantage of the cohesion graph (Trausan-Matu, Dascalu, & Dessus, 2012). The second is a novel approach that evaluates collaboration as an intertwining or overlap of voices pertaining to different speakers. The main difference between the two is that the first focuses on the ongoing conversations, therefore on its longitudinal dimension, whereas the later considers subsequent slices of the conversation, the synergy of voices, in other words the transversal dimension. By applying a greedy algorithm (Cormen et al., 2009) on both approaches, the overlap between the identified intense collaboration zones is remarkable.

9.2.1 Social Knowledge-Building Model

The actual information transfer through cohesive links from the cohesion graph obtains two valences by enforcing a personal and social knowledge-building process (Bereiter, 2002; Scardamalia, 2002; Stahl, 2006b) at utterance level. Firstly, a *personal dimension* emerges by considering utterances with the same speaker, therefore modeling an inner voice or continuation of the discourse. Secondly, inter-changed utterances having different speakers define a *social perspective* that models collaboration as a cumulative effect. Although similar to some extent to the gain-based collaboration model (Dascalu, Rebedea, et al., 2010; Trausan-Matu, Dascalu, & Rebedea, 2012), the transition towards Stahl's model of collaborative knowledge-building (see Figure 5) and the use of the multi-layered cohesion graph instead of the utterance graph are the main differentiators when addressing this computational knowledge-building model that enables a deeper and a more generalized analysis of collaboration in CSCL conversations.

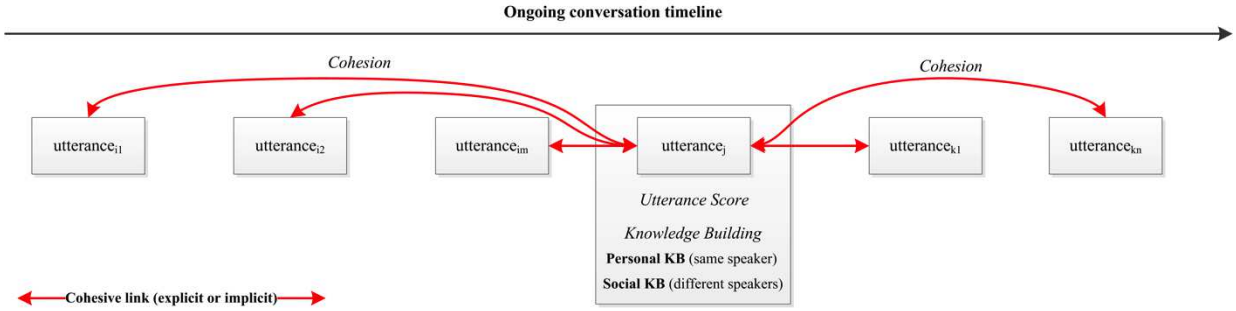


Figure 61. *ReaderBench* (3) Slice of the cohesion graph depicting inter-utterance cohesive links used to measure personal and social knowledge-building effects.

Therefore, each intervention or utterance now has its previously defined importance score (see *7.4* Cohesion-based Scoring Mechanism) and a knowledge-building (KB) effect, both personal and social (see Figure 61). The personal effect is initialized as the intervention's score, whereas the social effect is zero. Later on, by considering all the links from the cohesion graph, each dimension is correspondingly augmented: if the link is between utterances with the same speaker, the previously built knowledge (both personal and social) from the referred utterance is transferred through the cohesion function to the personal dimension of the current utterance; otherwise, if the pair of utterances is between different participants, the social knowledge-building dimension of the currently analyzed utterance is increased with the same amount of information (previous knowledge multiplied by the cohesion measure). In other words, continuation of ideas or explicitly referencing utterances of the same speaker builds an inner dialogue or personal knowledge, whereas the social

perspective measures the interaction with other participants, encourages sharing of ideas, fostering creativity for working in groups (Trausan-Matu, 2010b) and influencing the other participants' points of view during the discussion, thus enabling a truly collaborative discussion.

In this manner we can actually measure collaboration through the sum of social knowledge-building effects, starting from each intervention's score corroborated with the cohesion function. Moreover, personal knowledge-building addresses individual voices (participant voices or implicit/alien voices covering the same speaker), while social knowledge-building, derived from explicit dialog (that by definition is between at least two entities), sustains collaboration and highlights external voices. By referring to the dialogic model of discourse analysis, besides voices that are derived from the semantic chains in correlation to each participant's point of view, echoes are reflected by cohesion in terms of the information transferred between utterances, whereas the attenuation effect diminishes the strength of the cohesion link with the increase in distance between the analysis elements (see *7.2* Cohesion-based Discourse Analysis).

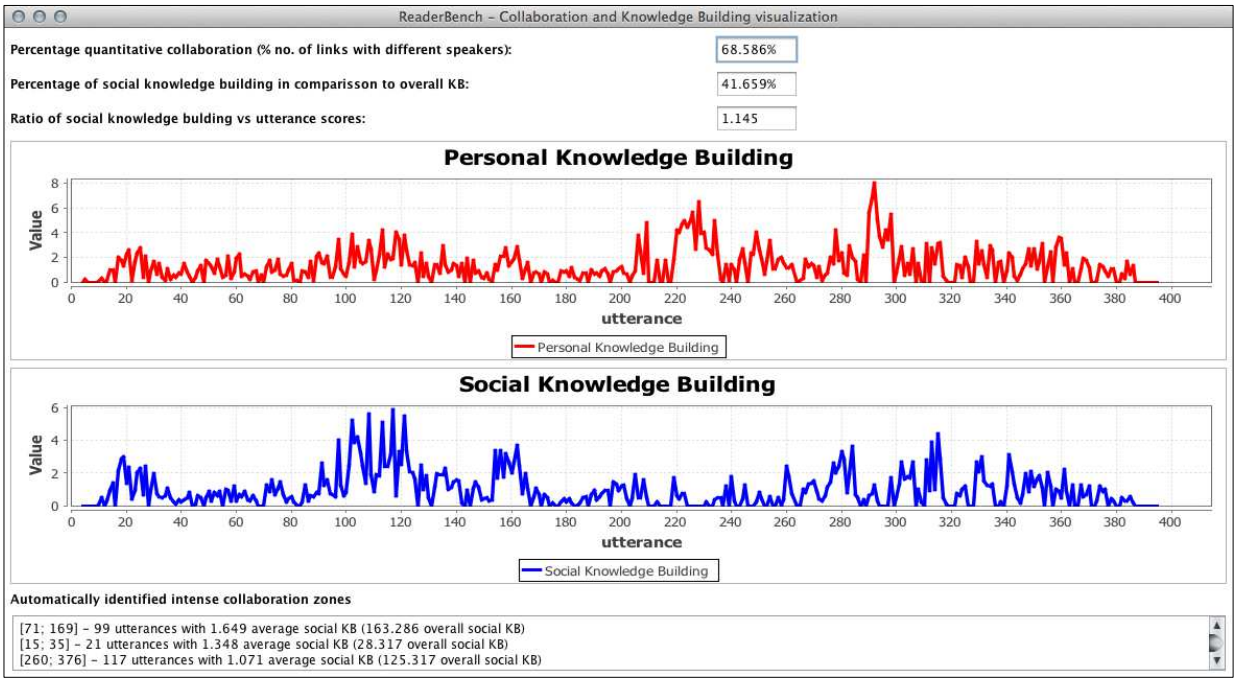


Figure 62. *ReaderBench* (3) Collaboration assessment and its evolution in time.
The interface introduces from top to bottom: a. the 3 overall collaboration factors as an overview of the conversation; b. individual graphs depicting the personal and social knowledge-building evolution throughout the entire discussion; c. the automatically identified intense collaboration zones with their corresponding span and cumulated social knowledge-building effect

Nevertheless, we must also consider the limitations of our implemented model in terms of personal knowledge-building. Collaboration clearly emerges from social knowledge transfer through cohesion

as the influence of one's intervention over other participants' discourse. In contrast, the approximation of personal knowledge-building rather represents an upper bound of the explicitly expressed information transfer between one's personal interventions. Similarly to the gain-based approach (Dascalu, Rebedea, et al., 2010; Trausan-Matu, Dascalu, & Rebedea, 2012), we use a quantifiable approximation of inner dialogue, although limited in terms of underlying cognitive processes. Personal knowledge-building is seen as a reflection of one's thoughts expressed explicitly within the ongoing conversation as cohesive links between interventions of the same chat participant. But this reflection does not necessarily induce personal knowledge-building, only a cohesive discourse. Therefore, we can consider that the computed value of personal knowledge-building is a *maximum value* of the explicit personal knowledge-building effect, modeled during the discourse through cohesive links.

In addition to the estimation of personal and social knowledge-building effects for each utterance and the modeling of their corresponding evolution throughout the conversation (see Figure 62), *ReaderBench* automatically identifies *intense collaboration zones* that are intervals of utterances in which participants are actively involved, collaborate and generate new ideas related to the ongoing context of the discussion. The first step within our greedy algorithm (Cormen et al., 2009) exploited in order to build up intense collaboration zones consists of identifying social knowledge-building peaks as maximum local values. Afterwards, each peak is expanded sideways within a predefined slack (experimentally set at 2.5% of the number of utterances); this slack was important due to our focus on the macro-level analysis of collaboration and due to the possible intertwining of multiple discussion threads. In the end, only zones above a minimum spread of 5 utterances are selected as intense collaboration zones.

In other words, after identifying the utterances with the greatest social knowledge-building effect, the algorithms expands each zone to the left and to the right, in a non-overlapping manner to previously identified zones, by considering utterances above the mean social knowledge-building value and that are in the previously defined slack. If in the end, the zone covers more than the specified minimum spread, it is considered an intense collaboration zone. From a different point of view and highly related to the process of identifying social knowledge-building, cohesion binds utterances within an intense collaboration zone in terms of on-topic relatedness.

From a holistic perspective addressing the conversation viewed as a whole, three factors were implemented in order to best characterize the overall collaboration within the discussion (see Figure 62). Firstly, *quantitative collaboration* is determined as the percentage of links from the cohesion graph having different speakers in comparison to the number of links automatically identified. Although rough as estimation, this measurement provides good insight with regards to the actual information exchange between participants. Secondly, the *overall social knowledge-building score* is compared to the overall knowledge-building effect. Thirdly, the *ratio* between the *overall social knowledge-building score* and the *overall utterance importance scores* is computed for highlighting the amount of information that is transferred through collaboration in comparison to what was withheld initially within each utterance.

9.2.2 Dialogical Voice Inter-Animation Model

In order to achieve genuine collaboration, the conversation must contain a dense intertwining of voices derived from key concepts and covering multiple participants of the conversation (Trausan-Matu & Rebedea, 2009; Trausan-Matu, in press). In order to obtain a computational model, a shift of perspective is required, from the voice synergy effect, towards the participant's point of view. As collaboration is centered on multiple participants, a split of each voice into multiple viewpoints pertaining to different participant is required (see Figure 63). A viewpoint consists of a link between the concepts pertaining to a voice and a participant, through their explicit use within one's interventions in the ongoing conversation. Moreover, we opted to present this split in terms of implicit (alien) voices (Trausan-Matu & Stahl, 2007), as the accumulation of voices through transitivity in inter-linked cohesive utterances clearly highlights the presence of alien voices. In addition, this split presentation of semantic chains per participant is useful for observing each speaker's coverage and distribution of dominant concepts throughout the discussion.

In addition, in order to identify the voice overlaps now pertaining to different participants, we changed from an ongoing longitudinal analysis of the discourse, presented in the previous section, to a transversal analysis of a context consisting of five adjacent utterances (with a possible shortening of the window, if the pause between adjacent utterances is greater than the imposed threshold) (see *7.5* Dialogism and Voice Inter-Animation). Subsequently, in order to evaluate collaboration following the conversation's timeline, we used a sliding window that models through its replication the overlap of voices pertaining to different participant in different contexts. More specifically, we

use a cumulated value of pointwise mutual information (PMI) obtained from all possible pairs of voices pertaining to different participants (different viewpoints), within subsequent contexts of the analysis (see Figure 64). In the end, a similar process of identifying intense collaboration zones based on the greedy algorithm described in the previous section is applied.

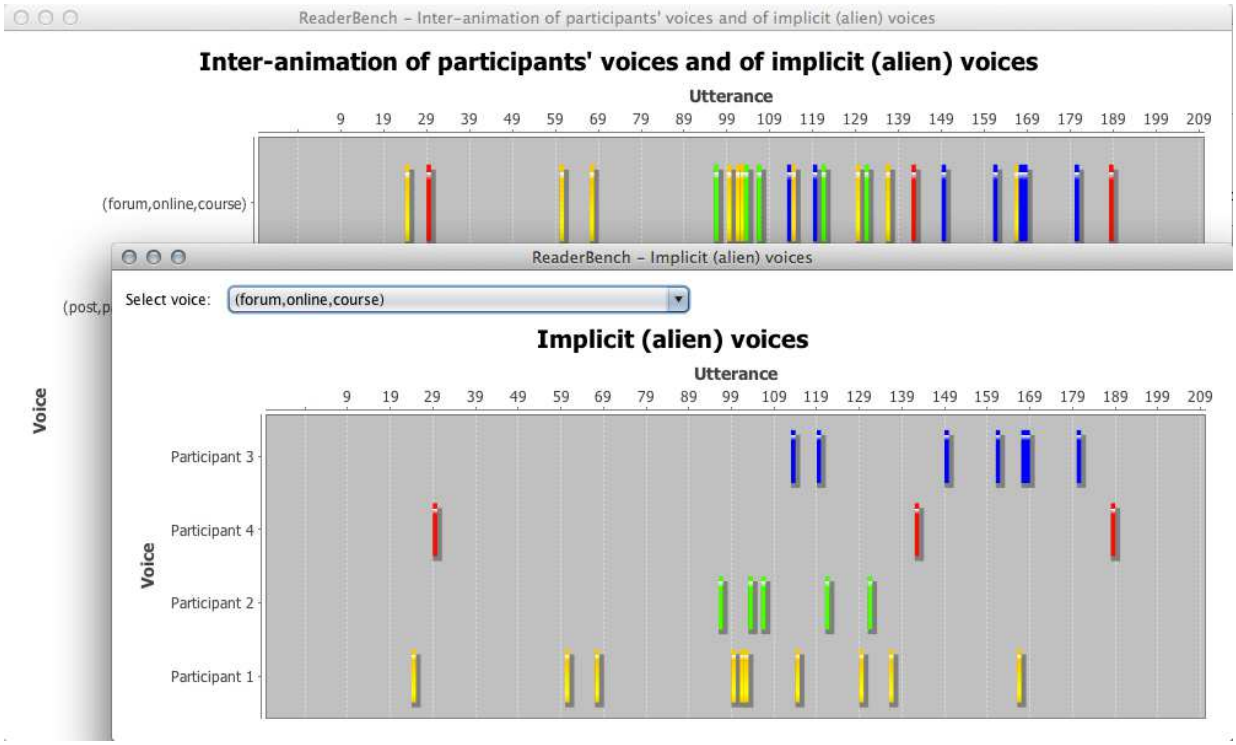


Figure 63. *ReaderBench* (3) Implicit (alien) voices split per participant and spread throughout the conversation.
The window frame from the background depicts the *(forum, online, course)* voice that was split per participant in order to highlight personal coverage of the conveyed concepts. The initial distribution of the voice can be obtained by overlapping the individual implicit (alien) voices for all participants

The inter-animation frame from Figure 64 presents the voices with the longest semantic chain span throughout the conversation. Each peak of collaboration obtained through PMI corresponds to a zone with a high transversal density of voices emitted by different speakers (e.g., around utterances with the following identifiers 110, 136, 225, 280 or 350). Two important aspects need to mentioned: 1/ as the algorithm uses the moving averages and applies PMI on sliding windows, the user must also consider a frame of 5 utterances in which each individual occurrence is equally dispersed (if not the case of a split horizon due to a pause in the conversation) and 2/ all the voices from the conversation are considered (even those that have as low as 3 constituent words); this explains greater cumulative values encountered in the graph (e.g., the excerpt centered on utterance 136 in which all conversation

participants are engaged and in which multiple concepts, pertaining to different voices, are encountered).

<Turn nickname="Participant 2"><Utterance genid="134" time="03.47.38" ref="130">wiki wiki means rapidly in Hawaiian language</Utterance></Turn>

<Turn nickname="Participant 3"><Utterance genid="135" time="03.48.31" ref="0">the forum was the place where in roman times people used to come and talk business</Utterance></Turn>

<Turn nickname="Participant 1"><Utterance genid="136" time="03.49.01" ref="135">and now the next best thing could be the blog – where someone shares its knowledge</Utterance></Turn>

<Turn nickname="Participant 2"><Utterance genid="137" time="03.49.22" ref="134">so it is a very quick way of letting others know what you have discovered</Utterance></Turn>

<Turn nickname="Participant 4"><Utterance genid="138" time="03.50.31" ref="136">yes, but knowledge is stored in books</Utterance>.

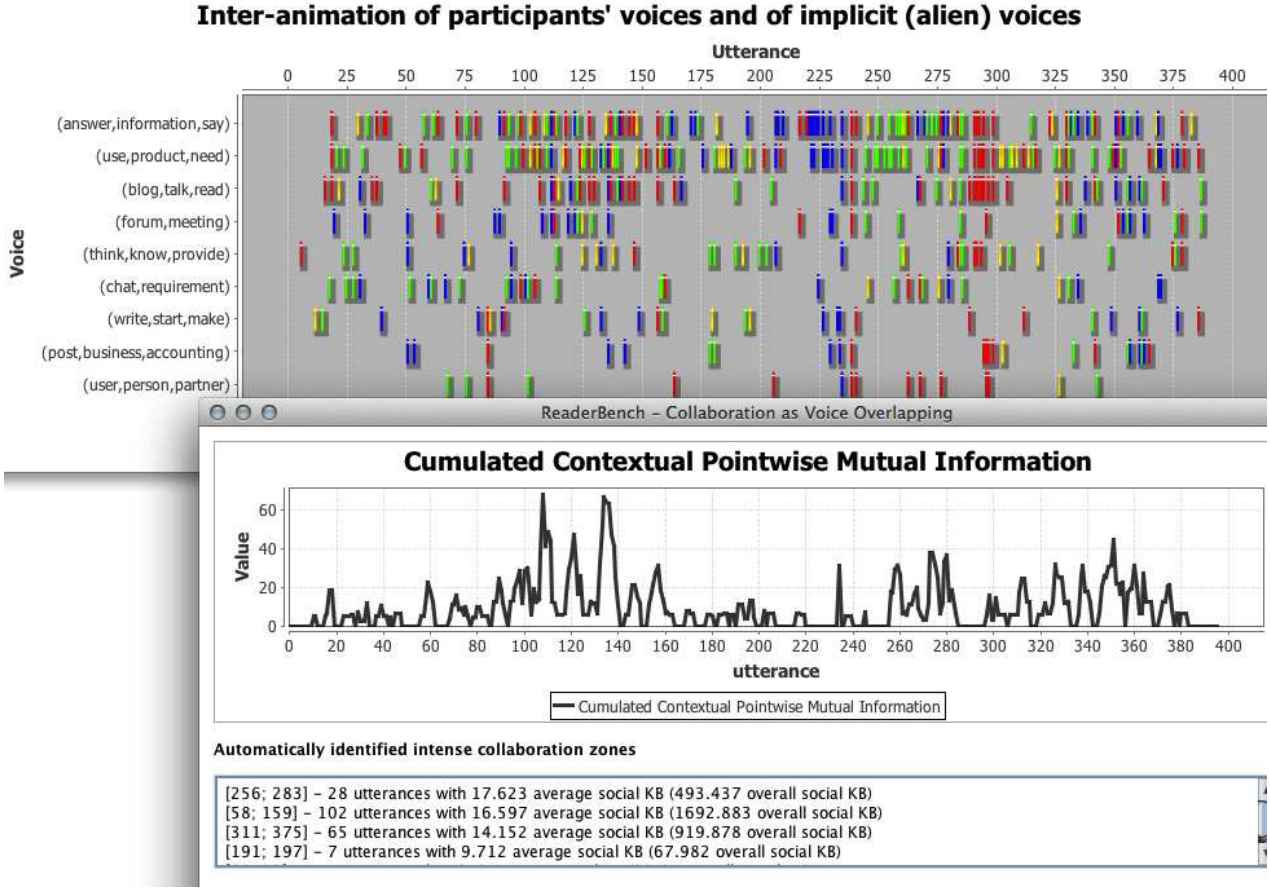


Figure 64. *ReaderBench* (3) Collaboration evolution viewed as voice overlaps between different participants (intertwining of different viewpoints), including the automatic identification of intense collaboration zones. The presented participant meaningful interventions denote a peak value in the collaboration evolution graph in the [134; 138] range where multiple voices pertaining to all conversation participants (e.g., knowledge, wiki, forum, blog, chat, people) co-occur

As an analogy, from an individual point of view, participant's overall collaboration can be seen as the cumulated mutual information between his viewpoints and all other participant viewpoints. In other

words, for a given participant, we compare through mutual information his viewpoint or individual voice distribution to all other speakers' viewpoints, for all voices identified in the conversation. Therefore, by comparing individual voice distributions that span throughout the discussion, collaboration emerges from the overlap of viewpoints pertaining to different participants.

9.2.3 Validation of Collaboration Assessment

Preliminary experiments (Dascalu, Trausan-Matu, et al., in press) were conducted in order to validate the dialogic models used for evaluating chat conversations, with emphasis on participant involvement and collaboration assessment. Three chat conversations conducted in an academic environment, with students from the 4th year undergoing the Human-Computer Interaction course and debating on CSCL technologies, were manually assessed by 4 tutors. More specifically, each student had to focus on a CSCL technology (chat, wiki, blog or forum), to present and debate on its benefits in specific use case scenarios generated throughout the conversation. These three conversations (Team 4, Team 34 and Team 36) were selected for detailed analysis after an overview of approximately 50 discussions engaging more than 200 students. Although high discrepancies were noticed in terms of the quality of the content, the involvement and the collaboration of its participants, these conversations were considered representative for the entire sample and the preliminary evaluations were conducted only on these conversations due to the high amount of time it takes to manually assess a single chat conversation (2 to 4 hours for a deep understanding of involvement and of collaboration).

Additionally, the time evolution interface depicted in Figure 65.a was developed in order to facilitate the manual evaluation of chats in terms of intense collaboration zones. In this context, the presentation of the conversation follows the timeline and models the intertwining of utterances, based on the cohesion graph. This component is useful for manually identifying: 1/ breaks within the conversation, zones with limited or no collaboration, due to the fact than within a specific time-frame we have a monologue of a participant, without any interventions from other users, and 2/ zones with high collaboration due to the dense inter-animation of utterances between different participants. In the particular case presented in Figure 65, all utterances with identifiers between 27 and 50 belong a single user, within a limited timeframe, therefore making the social knowledge-building effect zero. Afterwards, as multiple participants get involved in the ongoing discussion, collaboration increases.

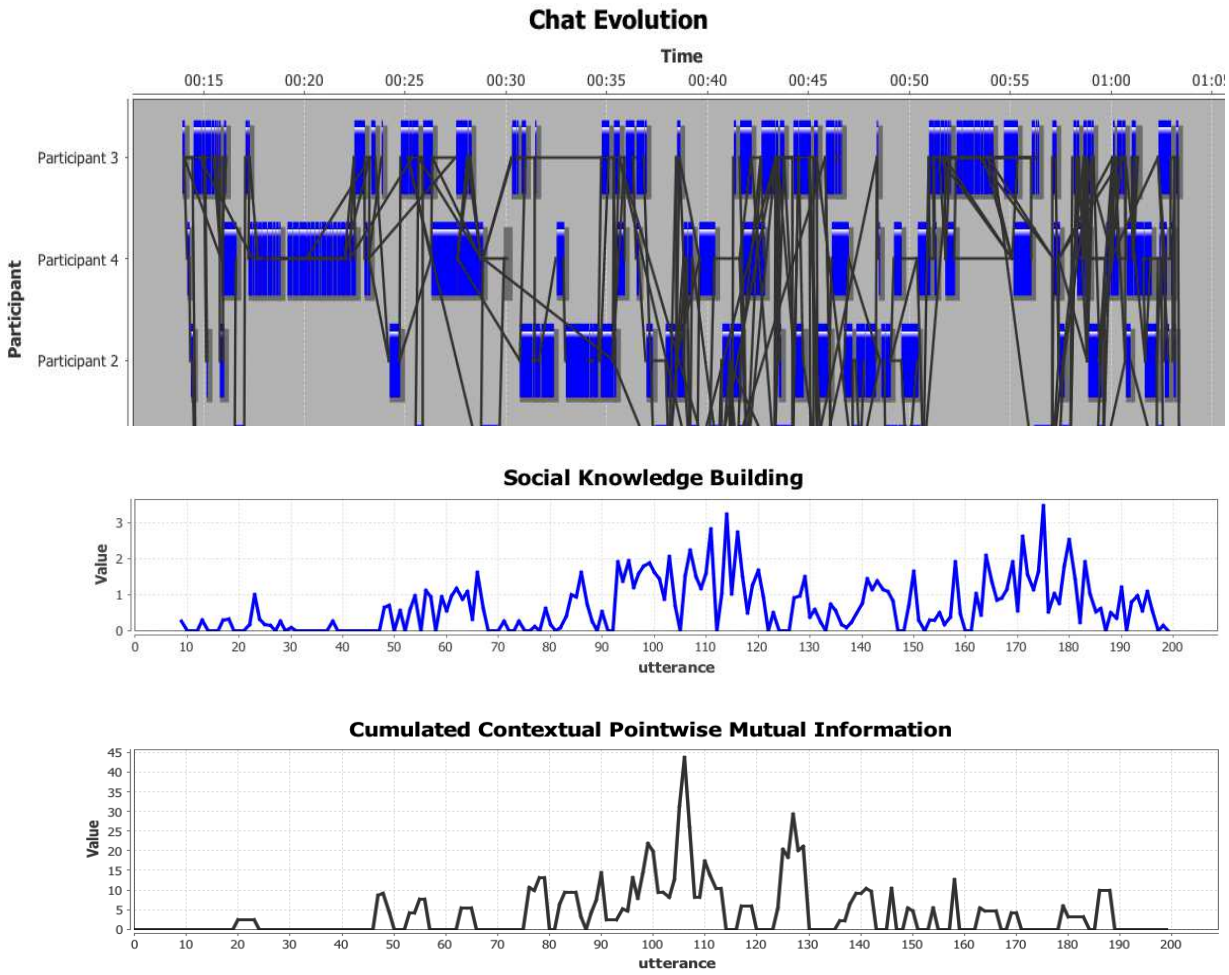


Figure 65. *ReaderBench* (3) Time slice of a conversation highlighting cohesion links and a monologue. Matching graphs of: a) Evolution in time of the chat conversation; b) Collaboration evolution seen as social knowledge-building; c) Collaboration evolution derived from voice overlapping

Table 31 presents the correlation between different evaluation factors extracted from *ReaderBench* and the final grades assigned by the experts. Although the participant's identifiers coincide, each conversation had different students attending it. Moreover, in order to ensure the equitability of our analysis, the correlations between the factors automatically determined by the system and the average values of the grades manually assigned by the experts were computed after combining the participants' scores from all conversations.

As an interpretation of the results presented in Table 31, we can observe in Team 4 conversation a discrepancy, as the involvement of the participants from a personal point of view was good, while the actual collaboration throughout the conversation is highly unbalanced. Team 34 conversation has the lowest scores in all the factors, whereas Team 36 conversation, that was considered the best by the tutors in terms of both quality and involvement, has the highest scores assigned by the system.

Table 31. *ReaderBench* (3) Correlation between manual and automatic participants' evaluations.

Participant Name	No. Utter.	Overall Utter. Score	Overall Personal KB	Overall Social KB	MI Viewpoint Overlap	Expert Grade 1	2	3	4	Avg.
				Team 4						
Participant 1	90	90.53	138.09	99.77	223.92	9.0	10.0	8.0	9.5	*9.13*
Participant 2	61	47.22	60.80	75.04	199.55	8.0	9.0	8.5	10.0	*8.88*
Participant 3	120	95.18	185.70	86.44	232.39	8.0	6.5	9.0	8.0	*7.88*
Participant 4	118	92.24	136.80	111.05	240.61	9.0	10.0	8.0	9.5	*9.13*
				Team 34						
Participant 1	23	21.43	35.34	22.77	82.20	6.0	6.0	7.0	6.0	*6.25*
Participant 2	34	25.02	32.69	25.39	105.30	5.0	7.0	7.0	6.0	*6.25*
Participant 3	73	44.83	74.80	44.72	100.03	8.0	7.0	8.0	7.0	*7.50*
Participant 4	60	45.38	85.41	33.68	110.57	7.0	4.0	7.0	5.5	*5.88*
				Team 36						
Participant 1	54	55.53	71.11	99.61	223.53	9.0		9.5	8.0	*8.83*
Participant 2	67	69.91	95.20	111.83	313.56	10.0		8.0	10.0	*9.33*
Participant 3	119	134.45	236.91	145.82	288.53	9.0		8.0	8.5	*8.50*
Participant 4	57	60.91	81.77	98.66	271.21	9.0		9.5	8.0	*8.83*
				Overall – all conversations						
Correlation	.64	.79	.69	.89	.84					

Moreover, by analyzing each factor's correlation, it becomes quite clear that the tutors emphasized on the quality of interventions, not on the mere number of utterances or their inter-dependencies. Additionally, the social knowledge-building dimension and the collaboration extracted from the mutual information of participant viewpoints are better correlated with the expert's grades; this sustains that collaboration was more important in the expert's evaluation than the personal effect of each participant, reflecting his/her involvement. As the Intraclass Correlation Coefficient (ICC) was .61 on single measures and .86 on average, results in terms of intervention scores (qualitative involvement evaluation) and social knowledge-building and mutual information between viewpoints (collaboration assessment at individual level) correlate extremely well with the average expert grades.

Table 32. *ReaderBench* (3) Overlap between manual and automatic identification of intense collaboration zones.

Conversation	Number of utterances	Manually annotated collaboration zones	Automatically identified intense collaboration zones	
			Social KB	Voice Overlap
Team 4	389	[90; 160] [320; 360]	[15; 35] [71; 169] [197; 208] [238; 245] [260; 376]	[33; 39] [58; 159] [191; 197] [256; 283] [311; 375]
Team 34	190	[90; 120] [170; 178]	[48; 66] [93; 121] [127; 134] [140; 150] [158; 184] [190; 195]	[47; 56] [76; 129] [138; 170]
Team 36	297	Relatively uniform distribution	[21; 126] [136; 182] [199; 241] [270; 288]	[18; 124] [139; 149] [188; 196] [205; 217] [249; 257] [271; 287]

In terms of intense collaboration zones, manual annotations and automatically identified zones are presented in Table 32, whereas the comparison between the zones identified through the two automatic collaboration assessment methods is covered in Table 33. The manual annotations were not covered in the later table as for Team 36 the tutors agreed that collaboration was uniformly distributed, thus making an automatic comparison inapplicable. Moreover, by analyzing the results from Table 32 and Table 33 we observe a good overlap in terms of accuracy measured as precision, recall and F-score (Manning et al., 2008) between the two computational models. This proves that one model is consistent with the other, but also a good match with the tutor annotations, therefore demonstrating the feasibility of our two approaches. The rather low correlation scores in Table 33 are completely justifiable as the two models are built from orthogonal dimensions of conversation analysis and consider completely different mechanics of evaluation, but in the end both properly address the purpose of identifying intense collaboration zones.

Table 33. *ReaderBench* (3) Overlap measurements between automatic models used to identify intense collaboration zones.

Conversation	Precision	Recall	F-score	Correlation
Team 4	.87	.71	.78	.41
Team 34	.68	.65	.67	.27
Team 36	.88	.67	.76	.48

Based on the previous analyses, the following indicators of bad collaboration (mostly in Team 34 conversation) were observed: 1/ the high number of automatically identified zones containing 1 to 3 utterances, which were not considered intense collaboration zones in the end, 2/ the low average value of social knowledge-building effect and 3/ no automatically identified collaboration zone with a wide spread (over 50 utterances, although it was the shortest conversation of the three). In contrast, conversations with good collaboration (namely Team 36 which had the best overall collaboration) have: 1/ higher average values of social knowledge-building and 2/ a more balanced distribution and higher coverage of the entire conversation in terms of the automatically identified intense collaboration zones.

Additionally, we have performed an evaluation for proving the inter-dependencies between the two collaboration assessment models: starting from all explicit links added by users in the chat environment (Holmer et al., 2006), we have measured the correlations between the cohesion scores and the similarities between utterances in terms of voice distributions; the later similarity is computed as a Pearson correlation between the utterances' voice occurrences. Linked with the nature of the evaluations (overlap of semantic chains versus an aggregated cohesion function), results were though medium: *average r* = .46, with *r*(Team 4 with 106 explicit links) = .54, *r*(Team 34 with 76 explicit links) = .48 and *r*(Team 36 with 226 explicit links) = .34.

Although the perspectives of the two collaboration assessment models are orthogonal while observing the unfolding of a conversation, there are multiple resemblances between the two proposed computational models. Firstly, the evaluation process of collaboration is based, in some extent, on the exchange of information between different participants; whereas in the first case, cohesion expresses the strength of the link in terms of the social knowledge-building effect between interventions of different chat participants, in the second voice overlaps are considered only while comparing different viewpoints and the exchange is expressed through mutual information.

Secondly, cohesion, seen as a link between analysis elements and an equivalent to a voice's echo, is caught in some degree through the process of overlapping occurrences of semantic chains, smoothed in predefined conversational contexts. Thirdly and most importantly, although one method is based on the effect of social knowledge-building and the other on the intertwining or overlap of voices belonging to different speakers, both computational models support dialogism and emphasize the dialogical perspective of collaboration in CSCL environments.

9.3 Long-term Discussion Groups Evaluation

Starting from the analysis of a single conversation, our aim in terms of assessing discussion groups consists of providing an automatic aggregation facility of multiple conversations, of building a global social network with all the involved participants and of verifying the validity of the automatic analysis proposed in *ReaderBench*, applied on a larger scale. Long-term discussion groups depict a set a participants or members of that group involved in subsequent conversations, over a longer timespan. This discrepancy between a local view, initially introduced in *A.S.A.P.* and *Ch.A.M.P.*, and continued by *PolyCAFe*, and a global one has multiple implications as specific technical aspects needed to be taken into consideration when merging a multitude of discussion threads. Therefore, from a technical perspective, the shift required a normalization of individual conversation scores, performing a distributed analysis due to the size of the corpus of discussion threads (Dascalu, Dobre, et al., 2011) (see *6.4.5* Distributed Computing Framework) and building a global interaction graph between all the participants. In the end, in order to perform the validation of the automatic importance scores, a critical thinking assessment framework was used to annotate the relevance of members' messages (Weltzer-Ward, Baltes, & Lynn, 2009).

Moreover, we specifically limited the perspective to long-term discussion groups, without clearly pinpointing at this moment the particularities of communities of practice (CoP) (Lave & Wenger, 1991; Wenger, 1999), as further refinements of our automatic assessment procedure are required to best fit the specificity of such communities. Nevertheless, the overall conducted study (Nistor & Fischer, 2012; Nistor et al., 2013; Nistor et al., in press) was positioned at the intersection of development of the expert status in CoP (Nistor, 2010; Nistor & Fischer, 2012) and technology acceptance (Bagozzi, 2007; Nistor, Lerche, Weinberger, Ceobanu, & Heymann, 2012). In other words, in terms of educational practice, the conducted study represented an extended application of

ReaderBench towards monitoring and assessing participation and collaboration (Strijbos, 2011) in communities.

The study included $N = 179$ participants (20 full-time faculty employees and 159 part-time faculty members), all of them holding a doctoral degree. The automatic analysis was focused on 3 variables extracted from all the messages of the asynchronous forum discussions (7370 interventions) available between August 2010 and June 2012: *participation*, *expertise* and *expert status* (Nistor et al., in press) (see Table 34). The intensity of participation was operationalized as the number of interventions of each group member; the quality of these interventions (utterance scores determined by *ReaderBench*) was considered an indicator of expertise; expert status was measured as in-degree and betweenness centrality within the group interaction graph (see Figure 66).

Table 34. *ReaderBench* (3) General long-term discussion group statistics ($N = 179$).

Factor	Mean	Standard deviation
1. *Participation*(number of interventions)	41.17	95.07
2. *Expertise* (quality of interventions or cumulative utterance importance scores)	160.53	418.94
3. *Expert status* (In-degree from interaction graph)	289.34	761.43
4. *Expert status* (Betweenness from interaction graph)	266.08	703.20

More specifically, through the dynamics of the interaction and through the quality of the interventions, a member of the discussion group obtains in time the status of expert. In addition, as most discussions followed a simple pattern – an inquiry, administrative or related to educational sciences, was initially formulated as a new thread and other members of the group responded subsequently -, participation can be considered mediated by the quality of the interventions, as only members with valuable insight contribute in each academic discussion thread (Nistor & Fischer, 2012). Moreover, participation influences the expert status determined after a longer timespan and reflected through specific SNA factors in a central position within the group. These dimensions of the analysis, with corresponding inter-dependencies, were later on studied in extent in (Nistor et al., 2013; Nistor et al., in press).

Although Table 34 depicts a high discrepancy across the group members in terms of involvement and of participation as the standard deviation values are approximately three times greater than the averages, these values are consistent for all analysis variables. Moreover, as expected because we are

dealing with a large group equitable between members, the cross correlations between the variables were high ($r > .70$). This can be also explained from the perspective that the analysis was focused on topics with a broad diversity and that the involvement of members, with similar backgrounds, within an academic environment, was balanced in terms of the impact of each intervention measured in the utterance's importance score.

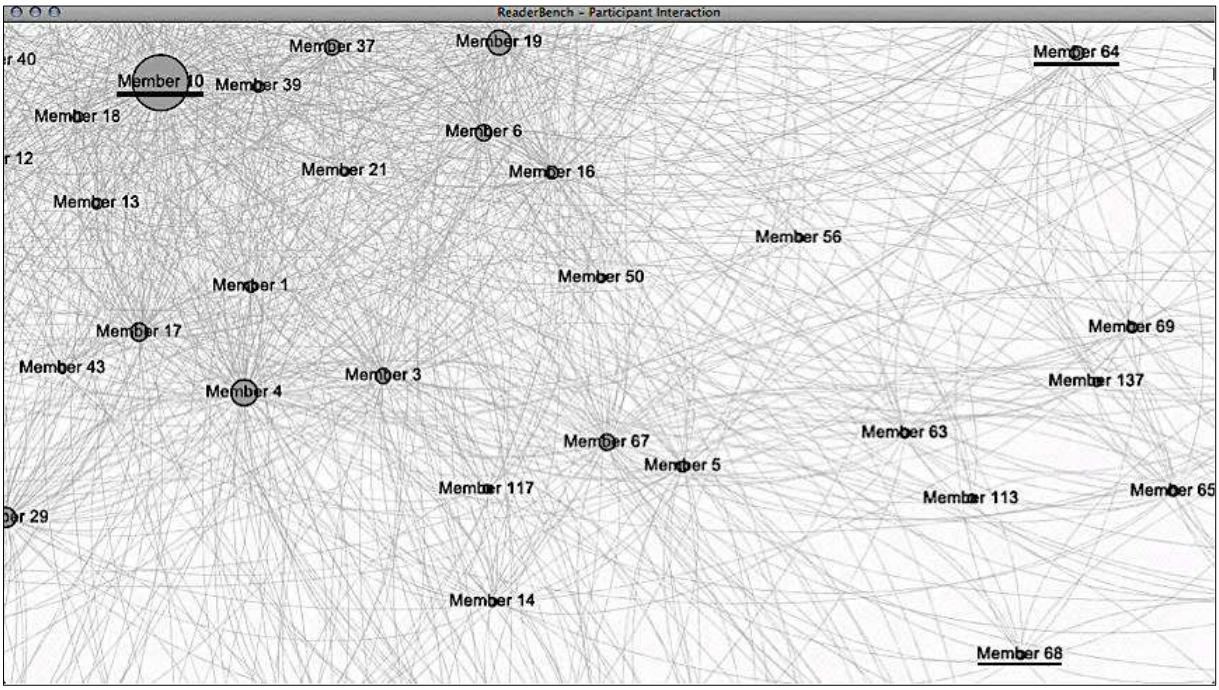


Figure 66. *ReaderBench* (3) Partial view of a group interaction graph.
A clear demarcation can be observed between different types of users: e.g., *Member 10*, by far the most actively involved member (830 interventions, in contrast to the second and third most active members: *Member 26* – 510 and *Member 19* – 458) and *Member 64* (68 interventions) or *Member 68* (12 interventions)

In contrast, if we analyze the average quality of each intervention per member (expertise divided by participation), there are rather high fluctuations for all members of the group ($M = 6.62$, $SD = 4.67$) and for the 10 most actively involved members in terms of participation ($M = 9.20$, $SD = 4.03$), with no correlation to any other variable. This allows us to consider that, although participation and expertise are highly correlated, this is a cumulative effect induced at group level, without any direct dependency between quantitative and qualitative evaluations of interventions. Nevertheless, the high correlation also resides in the intrinsic dependency that more interventions increase the overall importance score, as the scoring function (see *7.4* Cohesion-based Scoring Mechanism) always returns a positive result for each intervention.

Regarding the validation of the cohesion-based scoring mechanism, the bivariate correlation between the average relevance of messages determined manually (Weltzer-Ward et al., 2009) and the cumulative intervention scores per participant was of $r = 0.72$, $p < .001$ (for 414 messages sent by $N = 15$ discussion participants), which clearly demonstrates the adequacy of the scoring mechanism proposed in *ReaderBench*.

The visualization of the entire long-term discussion group was also of particular interest. Although the aggregated interaction graph uses the same measures described in section *9.1* Participant Involvement, the visualization became more relevant when applied on a larger scale, as the *expert status* now is visibly reflected in the dimension of each node (directly proportional to its betweenness score) and in a more central position within the social network graph (see Figure 66). All group member names have been anonymized to avoid privacy issues and the indexes are attributed in the order of first occurrence within the discussion group.

Additional experiments were conducted for splitting the discussions on two topics (research centered and administrative), therefore addressing the specificity of each intervention and focusing on the extraction of two sub-groups, hopefully as disjunctive as possible. List of concepts were manually built through questionnaires administered to 3 domain experts and included in the end 268 words for academic administration, respectively 857 words specific to educational sciences. Based on these lists, a new score of specificity was assigned to each analysis element, equal to its initial cohesion-based importance score multiplied by a normalized coverage of a given topic, seen as lemma overlapping between the predefined list and the words within each intervention. Therefore, besides the overall score that was initially assigned, each group member had a set of cumulative scores based on his/her interventions' specificity with regards to selected topics.

Table 35. *ReaderBench* (3) Statistics on the long-term discussion group specificity analysis ($N = 179$).

Evaluation scenario	$M_{specificity}$	$SD_{specificity}$
1. Administration	40.60	116.14
2. Educational sciences	33.53	90.15
3. Overall (equivalent to Expertise)	160.53	418.94

Starting from Table 35 and corroborated with the construction assumptions, we can conclude that: 1/ the group discussions were mostly administratively oriented as we obtained a greater average specificity by using a much shorter list of words; 2/ similar to the general scenarios, there was a high

variability in terms of expertise between the members, observable in the standard deviation approximately three times greater than the average value; 3/ the topics had a good cumulated coverage (46.18%) by using only 21.38% of the vocabulary/word lemmas mentioned throughout the group discussions; 4/ although the values suggest a rather clear categorization per topic, the final statistics for the sub-groups, each centered on a topic, have not induced a split of the initial long-term discussion group, but an overlap, as the majority of members addressed both topics throughout their interventions. This suggests a merge of the two topics by observing the entire interventions exchange during the long timespan, without a clear demarcation of membership to a sub-group.

9.4 Comparison of *ReaderBench* to *KSV*

Starting from the presentation of specific systems in *3.1.3* CSCL Computational Approaches, we considered the *Knowledge Space Visualizer* – *KSV* to be the most similar one to the current *ReaderBench* facilities (see Table 36). Both applications envision the visualization of participation and interaction between users through Social Network Analysis and semantic similarities between concepts or analysis elements, but the overall aims differ: 1/ *ReaderBench* is focused mostly on a deep analysis of each conversation/discussion thread with emphasis on involvement and collaboration, with the possibility of automatically aggregating them, whereas 2/ *KSV* was designed especially for obtaining an overview of interactions, with accent on visualization.

Table 36. *ReaderBench* versus *KSV* (Teplovs, 2008).

Benefits of *ReaderBench*	Benefits of *KSV*
Educational perspective	
Dialogical perspective induced by voice inter-animation	
Emphasis on collaboration in addition to a qualitative participation evaluation	A more shallow perspective of individuals and links between them
Conversation topics extraction relevant for highlighting the focus of the discussion	
The analysis is strictly based on textual information	Integration of addition relationships between 'notes' (e.g., annotation, authorial)

Benefits of *ReaderBench*	Benefits of *KSV*
Technical perspective	
Explicit or implicit/cohesive links between interventions are taken into consideration. Cohesion by itself integrates multiple perspective: semantic dimension, Latent Semantic Analysis and Latent Dirichlet Allocation	Multiple types of relations between the nodes are considered: structural (e.g., reply-to, build-on, reference, annotation, contains), authorial, or semantic (based on Latent Semantic Analysis)
Multiple NLP techniques applied on the initial interventions	
Post analysis centered on logs or excepts of conversations (chats and forum discussion threads)	Integration within the *Knowledge Forum* and the encouragement of continual analytic improvement
	Clustering of nodes
Aggregation of multiple discussion threads or conversations	Integration of multiple data sources
Comprehensive mechanism of utterance importance scoring	
	A multitude of parameters, configurations and visualizations available from the interface

10 Discussions

10.1 Advantages of our Approach

Starting from the integrated view presented in Figure 1, we have designed a cohesion graph (see *7.2* Cohesion-based Discourse Analysis) that was later on used to aggregate individual and collaborative learning through the underlying discourse structure. Without limiting the overall perspective, we opted here for focusing solely on *ReaderBench*, as it introduced the integrated cohesion-based analysis addressing both individual and collaborative learning, incorporating and refining nevertheless more functionalities than the other systems.

As particularities to demonstrate the intertwining of perspectives, we started from the cohesive properties of texts (Tapiero, 2007) and created a background for assessing group discussions. On the other hand, based on the polyphonic model designed initially for CSCL conversations, we were able to identify voices within general texts and observe the synergies among them. Moreover, comprehension was regarded from multiple perspectives: 1/ the identification of learner reading strategies, 2/ textual complexity, 3/ participant involvement assessment and 4/ collaboration assessment by enforcing the social knowledge-building model and the dialogical voice inter-animation model. In addition, as we are dealing with both individual and social learner activities, their intertwining has a higher impact on the outcome, as we are able to build in-depth scenarios, potentially more centered on creativity stimulation, more channeled and nevertheless flexible through the alternation of learner activities.

As readers make use of reading strategies when self-explaining, the automatic identification of such procedures plays an important role in the assessment of learner's comprehension. Through the proposed identification heuristics that make use of cohesion when addressing paraphrasing, bridging and knowledge inference, we are able to support tutors by providing a regularized and deterministic process, less prone to the subjective interpretations of the underlying strategies.

By combining different factors as readability, lexical and syntactic complexity, accuracy and fluency metrics, part of speech evaluation and characteristics of the parsing tree with cohesion expressed through lexical chains, information retrieval optimizations, semantic distances from *WordNet*, LSA and LDA, all embedded within the discourse model, we obtained an elaborated and multi-dimensional model, integrated in *ReaderBench*, capable of providing an overall balanced measure for textual complexity.

In terms of CSCL, starting from a dialogic model of discourse, our integrated analysis model can be used to identify cohesion gaps between utterances, to analyze participants' involvement and to evaluate collaboration individually and holistically, through the process of social knowledge-building or through voice inter-animation between different participants. The collaboration assessment model can be considered a cornerstone as it induces a dialogical approach and emphasizes the social knowledge-building perspective of collaboration in CSCL environments. In extent, by combining two different perceptions of cohesion, CSCL participants can use *ReaderBench* to assess to what extent utterance cohesion reflects group cohesion, as an outcome of collaboration depicted from the interaction graph.

Furthermore, based on the results of our validations, we can extrapolate that *ReaderBench* can be used to support tutors and teachers in: 1/ the cumbersome process of identifying reading strategies; 2/ the prediction of a complexity class for a reading material based on a wide variety of factors; 3/ the time-consuming process of manually assessing chat conversations. From a different computational point of view, the distributed architecture initial integrated in *PolyCAFe* (see *6.4.5* Distributed Computing Framework) also played an important role in terms of speedup and reliability of the *ReaderBench* platform. As proved, our parallel architecture performs and scales well under a wide variety of conditions and loads.

10.2 Faced Problems and Provided Solutions

While considering the wide variety and complexity of integrated functionalities, various problems were encountered in the design, implementation and validation phases of *ReaderBench*. Firstly, in terms of semantic models, the biggest problem was the corpus adequacy and specificity. As we started from general purpose, well renowned corpora (e.g., TASA or "Le Monde"), the corresponding LSA and LDA representations proved less efficient and more prone to noise when addressing certain

documents or conversations focused on a very specific domain or topic. Moreover, the performance of the models was affected for French, as we had to use a limited set of resources in terms of the Natural Language Processing pipe for French language.

In terms of reading strategies, higher-level strategies (bridging and knowledge inference) were hard to model as their informal definitions were far from any computational method or analogy. Therefore, multiple runs with different estimation functions had to be performed in order to choose the best one. Moreover, the selection of the threshold values used for semantic similarities showed a high fluctuation for each verbalization by itself, making the process even more difficult as we were dealing with a limited corpus – in other words, we faced the classic problem of over-fitting versus generalization.

While addressing textual complexity, the biggest problem we encountered was due to the lack of appropriate corpora for SVM training. Initial we used freely available online texts (Dascalu et al., 2012), but the collection was rather small and unreliable as no clear pattern of complexity emerged. Later we shifted towards predicting the class complexity based on DRP-derived classes (Dascalu, Dessus, et al., in press; McNamara et al., in press). Results significantly improved, but this is partially an artificial increase due the regularization of class assignments, as the DRP score is based on surface factors only. This proved that the model comprising all factors, ranging from surface to semantics, has an excellent overlap with DRP, but that morphology and semantics, considered separately from lower level factors, can be also used to accurately predict textual complexity, therefore highlighting the interdependencies between cohesion and complexity. In order to fine-tune the results, additional investigations and experiments are to be conducted to find the best parameters for the SVM, making predictions more reliable, whereas additional cohesion measurements will be performed for enriching the semantic perspective of our analysis.

In terms of chat conversations, besides surface problems consisting of spelling errors, elliptical expressions, jargon or inappropriate syntax, we faced two major problems: 1/ the subjectivity of the analysis, as the intertwining of multiple discussion threads in multi-participant conversations denatures the perception of each learner's involvement in longer discussion; 2/ the use of a small validation corpus of chats in the end as we were dealing with a time-consuming process.

Lastly, from a technical perspective, the necessary computing power and resources have become quite an issue (e.g., POS tagging and parsing for French can use more than 4GB of RAM on complex

sentences), plus the tasks for which a distributed, fault-tolerant, computing alternative became mandatory (e.g., for training semantic models using large corpora or for evaluating discussion groups consisting of multiple conversations). In this context, the use of distributed computing facilities has become a necessity, greatly impacting the overall performance of the learning platform.

10.3 Educational Implications

As the overall presentation of the systems (especially for *ReaderBench*) was centered more on comprehension and cognitive validations, in this section we strive at establishing a clear link to learning through the following: 1/ the *presentation of envisioned educational scenarios* that can be conducted in a classroom context, where learners effectively use *ReaderBench*, or in distance learning, where *ReaderBench* can be also integrated in the Learning Management System (LMS) via specific APIs (Application Programming Interfaces); 2/ the *integration* of *ReaderBench* as a *learning analytics* (LA) tool *into* the *learning design* process (Mor & Craft, 2012), more specifically into the 'Scenario Design Process Model' proposed by Emin, Pernin, and Guéraud (2009); 3/ the presentation of various pedagogical scenarios making full use of *ReaderBench*'s transferability, integrating also its extension towards the assessment of virtual Communities of Practice (vCoP) and of Communities of Inquiry (CoI).

Table 37. *ReaderBench* facility coding.

Code	*ReaderBench* facility	Section with detailed description
	Individual Learning	
RB.1	Generic Document Assessment	*7*
RB.1.1	Building the Cohesion Graph	*0*
RB.1.2	Topics Modeling and Topics Map Generation, including similar/inferred concepts	*7.3*
RB.1.3	Identification of Most Important Sentences	*7.4*
RB.1.4	Voices Identification and Inter-animation, spanning throughout the document	*7.5*
RB.2	Reading Strategies Assessment	*8.1*
RB.3	Textual Complexity Assessment	*8.2*

Discussions

Code	*ReaderBench* facility	Section with detailed description
	Collaborative Learning	
RB.4	Participant Involvement Assessment	*9.1*
RB.5.1	Collaboration Assessment based on the Social Knowledge-building Model (Individual Participant Evaluation and Conversation Evolution)	*9.2.1*
RB.5.2	Collaboration Assessment based on the Voice Inter-Animation Model (only Conversation Evolution derived from Alien Voices Inter-Animation)	*9.2.2*
RB.6	Participation, Expertise and Expert Status Evaluation in Long-term Discussion Groups	*9.3*

We opted to emphasize in this section the learning dimension solely by exploiting the features of *ReaderBench*, as it comprises more functionalities than the other systems, addressing both individual and collaborative learning, after refining all previous processes. Nevertheless, similar educational implications have been presented also for *PolyCAFe* in *6.2* Theoretical Considerations and Educational Scenario and *6.6* Conclusions and Transferability. Moreover, in order to ensure traceability between the previous described components of *ReaderBench* and the activities suggested within the following subsections, we defined a corresponding code for each major facility (see Table 37).

10.3.1 Envisioned Educational Scenarios

ReaderBench can be used as a Personal Learning Environment (PLE), allowing three kinds of work-loops in terms of individual learning (see Figure 67), in which teacher/learners can be freely involved (Zampa & Dessus, 2012). The first one is a *reading loop* in which learners read some materials (e.g., course text, narrative) and can, at any moment, get information about its textual organization from *ReaderBench*. The second one is a *gist selection loop*, a bit more interactive than the previous, that needs to be implemented. Learners produce keywords or select main sentences of the read texts and submit their selection to *ReaderBench*, which prompts feedback. The third is a *writing loop* that gives learners the opportunity to develop at length what they understood from the text (e.g., summaries) or the way they understood it (reading strategies applied on self-explanations). Besides these three loops, the tutor can use *ReaderBench* to select appropriate textual materials according to the learners' level.

Discussions

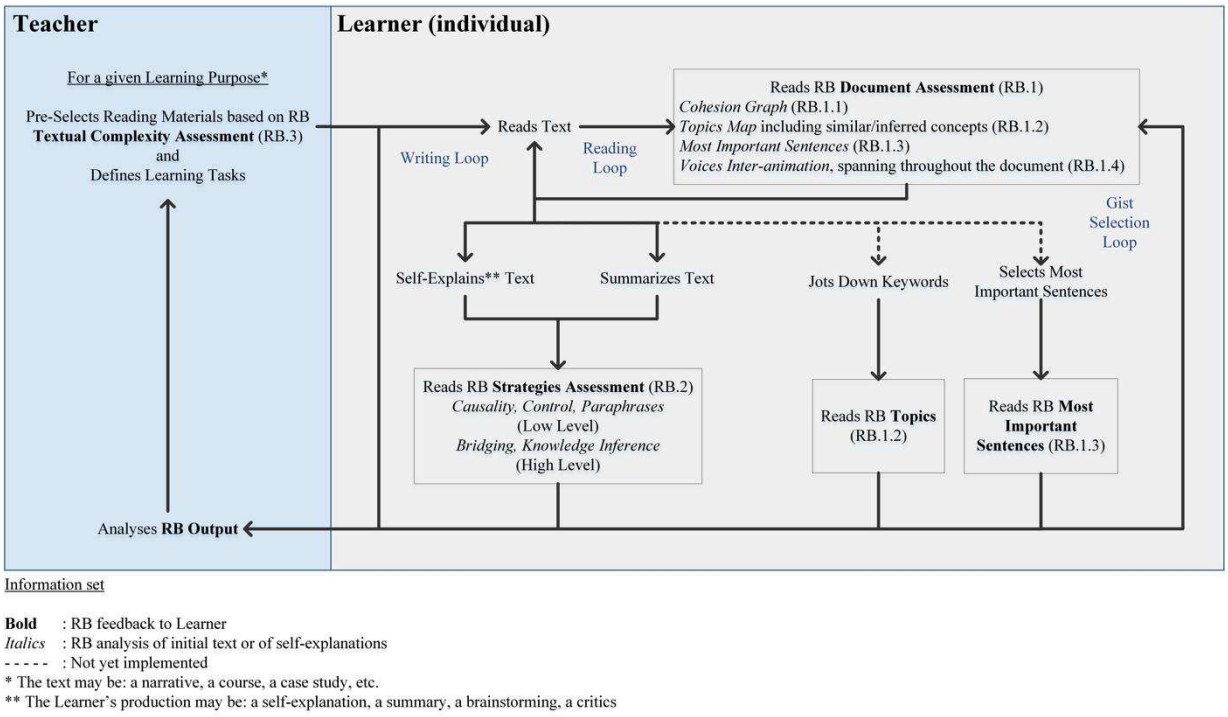


Figure 67. *ReaderBench* Educational scenario centered on individual learning and focused on the learner perspective.

In terms of collaborative learning, *ReaderBench* can be used as support for both learners and tutors enabling two types work-loops (see Figure 68). Within the *reading loop* learners are focused on other peer's interventions, as well as on an overview of the entire conversation (chats or forum discussion threads). The *writing loop* considers the learner's productions in the ongoing conversations, with emphasis on his/her participation and collaboration. The latter two elements can be also automatically assessed via *ReaderBench's* facilities that provides support for monitoring the evolution of participants' involvement, as well as for collaboration assessment, modeled through social knowledge-building and voice inter-animation. In addition, tutors can pre-select and assign learning materials to learners with an appropriate textual complexity level. Of course, collaborative processes can be alternated with individual learning scenarios in order to create more complex educational scenarios, applicable in both classroom and distance learning contexts, although customizations need to be taken into consideration.

Discussions

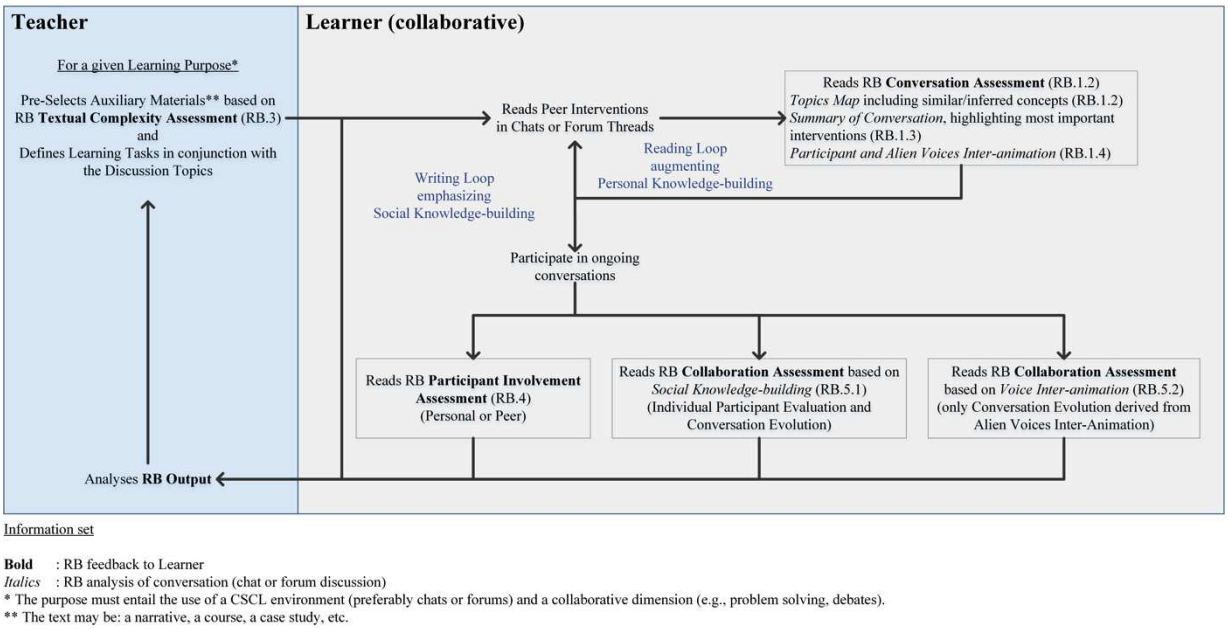


Figure 68. *ReaderBench* Educational scenario centered on collaborative learning and focused on the learner perspective.

In addition, from a different point of view, *ReaderBench* can be also used to *devise* a wide variety of learner *exercises*, ranging from voice annotation, topics identification, cohesion assessment between adjacent paragraphs, to the annotation of important sentences, later to be automatically evaluated in relation to *ReaderBench*'s outputs. All the previous scenarios are theoretical and need to be tested and validated in educational situations taking place in controlled environments, either classroom or distance learning specific.

10.3.2 Shifting the Perspective towards the Scenario Design Process Model

As the previous subsection was centered on learners, we thought it opportunistic to change the perspective and focus on teacher inquiry. Moreover, through the provided facilities, *ReaderBench* can be considered a post-analysis Learning Analytics (LA) tool (Cooper, 2012). Therefore, the transition towards learning design (Mor & Craft, 2012) and the integration of *ReaderBench* in the learning lifecycle became more than feasible. As various learning design cycles exist, we considered the "Scenario Design Process" model (Emin et al., 2009; Emin et al., submitted) the best fit due to its high degree of generality and its appropriates in terms of sample of pupils, ranging from 11 to 18 year old, similar to the ones that were involved in the experiments conducted within the ANR DEVCOMP project. The model was co-designed with groups of teacher-designers in the French

secondary educational system during the CAUSA project coordinated by the French Institute of Education (2005-2009).

The goal was to model the steps followed by a teacher in everyday practices while designing and using a learning scenario. The iterative lifecycle, depicted in the left-hand side of Figure 69, consists of three main steps: design, enactment and evaluation, with a perspective to capitalize and reuse the scenario again, and relies on an empirical study performed in two steps: firstly, the elicitation of the design process from two expert teachers and, secondly, the validation of this process by several groups of teachers.

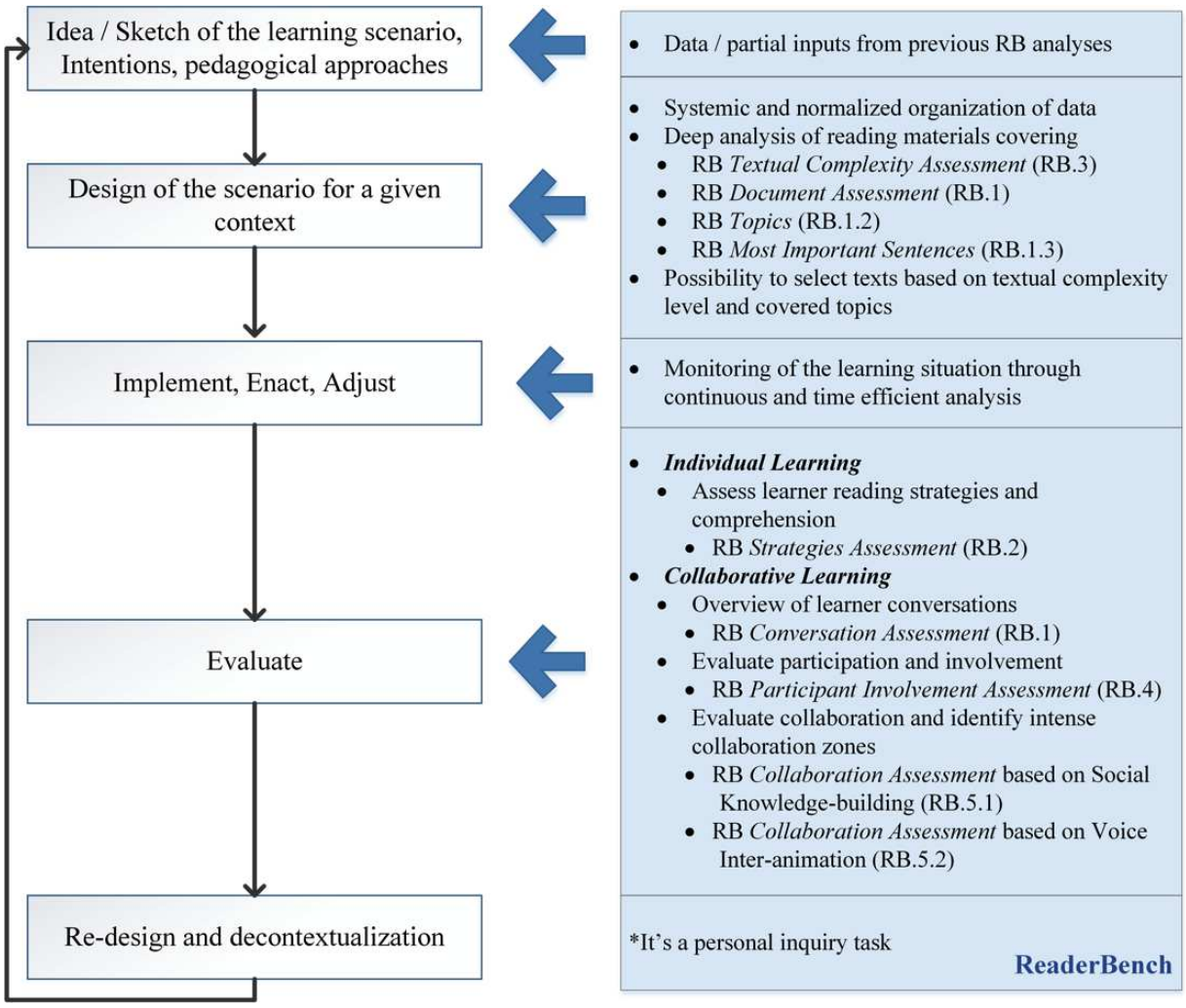


Figure 69. *ReaderBench* Integration as a Learning Analytics tool in the Scenario Design Process Model.

While following the major phases in the design of a pedagogical scenario by a teacher-designer, the first step consists of defining the *intentions* (in terms of learning outcomes, competencies and knowledge) and the *pedagogical approach* (e.g., the way of teaching, the role of the teacher). In this context, data from previous *ReaderBench* analyses could provide a strong incentive for a change. After

obtaining a general sketch/idea of the learning scenario, the actual *design* of the *class scenario* can begin. As a teacher integrates iteratively and progressively the different constraints of his/her specific context (domain-related, pedagogical, situational, administrative) (Emin, Pernin, Prieur, & Sanchez, 2007), s/he can select texts based on the complexity level indicated by *ReaderBench* and can consider each of its facilities to be integrated or not within the learner tasks.

The next step in the model assumes the implementation of the previously defined scenario or *enactment*, where the teacher adapts the scenario and achieves a different, 'on the fly', orchestration than the one s/he initially envisioned and designed. During this step, *ReaderBench* can provide monitoring facilities of the learning situation through continuous and time-efficient analyses that can be automatically conducted on the learner productions. After the actual implementation, the teacher *evaluates* the scenario and its successive adjustments. *ReaderBench* can provide decisive inputs at this stage as it can be used to automatically assess reading strategies and comprehension in terms of individual learning, as well as participation and collaboration of learners in CSCL environments, from a collaborative learning point of view. In the end, the overall evaluation enables *re-design*, comments on the scenario for further use, and a step of *decontextualization*, the definition of a 'pattern scenario' in order to share it with other teachers or to reuse it in another context.

10.3.3 Pedagogical Scenarios involving *ReaderBench*'s Transferability

In contrast to the previous subsection that envisioned *ReaderBench* as a tool in the design of a single educational scenario, this section offers a more general and tabular view of the integration of *ReaderBench* in already existing educational scenarios. Therefore, the analysis focuses on: 1/ the source texts that are considered as input, 2/ the activities or tasks for learners that need to be undergone and 3/ the resulting teacher/tutor activities for supporting learners through *ReaderBench*'s facilities. Within this scenarization process, Table 38 is centered on individual learning, whereas Table 39 explores the opportunities of supporting collaborative learning.

Table 38. Main pedagogical scenarios centered on individual learning and involving *ReaderBench*'s transferability.

Pedagogical Scenario	Source Text	Learner Activity (in *ReaderBench*)	Tutor/Teacher Activity (in *ReaderBench*)
Reading materials complexity evaluation	General texts and/or corpora of documents	-	Estimate textual complexity level for selection purposes (RB.3, RB.1)
Topics modeling	Source texts and peer's productions	Identify keywords (RB.1.2)	Annotate keywords and inferred concepts (RB.1.2)
Dialogism and voices highlighting	Source text and personal and peer's productions	-	Identify salient voices and voice inter-animation patterns (RB.1.4)
Understanding a given material/ Comprehension evaluation through reading strategies	General texts (e.g., several portfolio, cases studies)	Self-explain (RB.2)	Identify learner reading strategies for comprehension analysis (RB.2)
Summary production	Texts on a given domain	Write summary	Evaluate summary inner cohesion and cohesion with the initial texts, as well as coverage of important sentences and concepts (RB.1.1, RB.1.2, RB.1.3)
Lecture Notes Analysis or Learning academic writing	Several documents (e.g., tutorials)	Self-explain (RB.2) and write synthesis	Identify reading strategies (RB.2) Evaluate synthesis in terms of cohesion with initial materials (RB.1.1) Evaluate similarity of textual complexity levels with assigned texts (RB.3)

In terms of collaborative learning, of particular interest is *ReaderBench's* support, starting from the facilities presented in *9.3* Long-term Discussion Groups Evaluation, in terms of virtual Communities of Practice (vCoP) and Communities of Inquiry (CoI) that have shown an increasing interest lately. Shortly put, Communities of Practice (CoP) are groups of people sharing goals, activities, and experiences in the frame of a given practice (Lave & Wenger, 1991; Wenger, 1999). The community practice continues over lengthy periods of time and its termination is often neither planned, nor foreseen. Numerous communities are found in schools, universities and research institutes, either in face-to-face or in computer-mediated settings (Kienle & Wessner, 2006; Nistor & Fischer, 2012).

Participation in a CoP leads to the accumulation of experience, stimulates the social construction of knowledge and the development of expertise (Paavola, Lipponen, & Hakkarainen, 2004), hence, making it particularly interesting for educational research and practice.

Overall, CoPs are effective environments of knowledge sharing and knowledge creation (Paavola et al., 2004; Nistor & Fischer, 2012), therefore participation in CoP is desirable for many academic activities. Participation can be mediated by communication technologies, e.g. when CoP are geographically distributed, building thus the so-called virtual CoP (vCoP). In such environments, it would be of great advantage to assess participation using automated procedures. Previously, this has been done for collaborative learning in computer-mediated small groups (Strijbos, 2011); in vCoP, however, the basis for automated procedures is scarce due to insufficient quantitative evidences. Moreover, previous research has shown that a CoP member's expert status can be measured through SNA, determining a member's so-called centrality (Borgatti, Mehra, Brass, & Labianca, 2009).

In this context, the main *ReaderBench* facilities presented in Chapter *9 ReaderBench* (3) – Involvement and Collaboration Assessment through Cohesion and Dialogism, with emphasis on the customizations already performed in Section *9.3* Long-term Discussion Groups Evaluation, become more than appropriate to be used for evaluating participation, expertise and expert status (Nistor, Baltes, et al., submitted; Nistor, Dascalu, et al., submitted). Nevertheless, future refinements of the automatic methods need to be included in order to address the specificities of vCoP.

On the other hand, Communities of Inquiry (Garrison, Anderson, & Archer, 2000) "emerged in the specific context of computer conferencing in higher education—i.e., asynchronous, text-based group discussions—rather than from a traditional distance education theoretical perspective assumed that students worked independently from each other" (Garrison, Anderson, & Archer, 2010, p. 5). Starting from the three major elements or 'presences' within the framework (social, cognitive and teaching presence), *ReaderBench* can be used to support and evaluate certain presences or intersections among them (see Table 39). Overall, the possibility to integrate *ReaderBench* in vCoP and CoI, with focus on monitoring and evaluation, emphasizes its *flexibility* as a learning platform, through its applicability in various educational scenarios and its integration in various learner and tutor activities.

Table 39. Main pedagogical scenarios centered on collaborative learning and involving *ReaderBench*'s transferability.

Pedagogical Scenario	Source Text	Learner Activity (in *ReaderBench*)	Tutor Activity (in *ReaderBench*)
Collaborative problem solving	Mostly chats or forum discussion threads	Participate in the conversation Evaluate personal and peer involvement (RB.4) Assess personal and peer collaboration (RB.5.1, RB.5.2)	Provide external guidance (optional) Evaluate and rank participants (RB.4) Assess collaboration and identify intense collaboration zones (RB.5.1, RB.5.2)
Brainstorming	Chats	Participate in the brainstorming session Identify key concepts and ideas (RB.1.2) (personal or peer) Estimate cohesion to other peer's ideas (RB.1.1)	Mediate brainstorming session (optional) Identify key concepts and ideas (RB.1.2) (overall or individual participant) Evaluate participant involvement (RB.4) Assess collaboration in terms of intertwining ideas (RB.5.1, RB.5.2) Analyze cohesion between generated ideas (RB.1.1)
Debate	Chats or forum discussion threads	Participate in the debate Identify key concepts and viewpoints (RB.1.2, RB.1.4) (personal or peer)	Moderate debate (optional) Identify key concepts and viewpoints (RB.1.2, RB.1.4) (overall or individual debater) Analyze viewpoints cohesiveness and on-topic relevance (RB.1.1, RB.1.2) Identify intense/flaming regions of the conversation (RB.5.1, RB.5.2) Evaluate overall debater involvement and impact (RB.4, RB.6)

Discussions

Pedagogical Scenario	Source Text	Learner Activity (in *ReaderBench*)	Tutor Activity (in *ReaderBench*)
Virtual Communities of Practice	Forum discussion threads spanning	Participate in the community Share information and experiences Formulate and address inquiries Evaluate personal expertise and participation (RB.4, RB.6)	Mediate participation (optional) Share information and experiences Respond to inquiries Evaluation participation, expertise and expert status (RB.4, RB.6)
Community of Inquiry	Inquiry conversations	Generate notes or intervene during the integration and the resolution phases	Mediate interventions during resolution phases Evaluate *social presence* – Participation and involvement assessment (RB.6) Evaluate *cognitive presence* – Expertise and expert status assessment (RB.4, RB.6) *Support discourse* – Cohesion-based discourse analysis and topics modeling (RB.1.1, RB.1.2) *Select Content* – Textual complexity assessment (RB.3) Measure *inter-subjective agreement* through cohesion (RB.1.1)

11 Conclusions

11.1 Personal Contributions

The development of multiple systems, a constant growth in terms of the complexity of the approach, the multitude of considered factors, the unified approach that addresses both general texts and conversations and the emphasis on providing effective support for tutors and students in their learning and CSCL activities, are just the highlighting points of our research. We have provided an inter-disciplinary approach covering *informatics*, with emphasis on natural language processing as support for conducting all automatic assessments, *cognitive psychology*, in terms of validations and comprehension modeling with regards to cohesion, textual complexity and partially coherence, *educational sciences* in terms of implications, transferability and potential educational scenarios and *philosophy* while addressing Bakhtin's dialogism and polyphony.

Out of the four presented systems, *A.S.A.P.*, *Ch.A.M.P.* and *ReaderBench* were developed entirely by the author of this thesis, under the direct guidance of both thesis supervisors: Ștefan Trăușan-Matu and Philippe Dessus. In terms of *PolyCAFe*, as it was developed within the larger context of the European FP7 project LTfLL – Language Technologies for Lifelong Learning –, three central players were involved: Ștefan Trăușan-Matu, supervisor and coordinator of our team throughout the entire project, Traian Rebedea, responsible for the user interface and for coordinating the validation process, and Mihai Dascălu in charge of core functionalities development and application logics; therefore we found it appropriate to include the major findings of *PolyCAFe* in this thesis. Moreover, while addressing *ReaderBench*, the larger context in which the pupil's verbalizations were gathered and manually annotated is represented by the ANR DEVCOMP project – *Développement de Compétences* –, in which Philippe Dessus, Maryse Bianco (also coordinator of the project) and Aurélie Nardy were also actively involved in the design of *ReaderBench*.

Conclusions

ReaderBench can be considered the most advanced of our developed systems, a complex environment integrating new ways to assess a wide range of cognitive processes involved in reading and collaborating through the use of advanced NLP techniques. It provides a semantic insight and cohesion-based discourse structure through the combination of multiple semantic distances. Moreover, the design of *ReaderBench* considers two dimensions. On one hand, the *flexibility* of the environment is highlighted through the following features: comparison of complexity levels of several texts, one to another, and the ease of editing reading materials from within *ReaderBench*, with the possibility to also add dynamic breakpoints for learners' verbalizations or summaries. Teachers can thus *manipulate* textual materials in order to reach desired features. Also learners can very quickly have an idea of the way they regulate their reading (strategies assessment) or involvement in terms of collaboration (social knowledge-building and voice inter-animation collaboration assessment models). On the other hand, *extensibility* is reflected in the ease of training and of using additional LSA semantic vector spaces or LDA topic models or in the possibility to augment the features used for assessing textual complexity.

Of particular interest were the conducted *validations*, starting from simple score matching performed in *A.S.A.P.* and *Ch.A.M.P.*, followed by *in situ* validations in *PolyCAFe* mostly focused on learner and tutor feedback and continued by the thorough cognitive validations in *ReaderBench*, centered on providing a comparison to learners' performances. As for *PolyCAFe* two validation experiments and a participant ranking verification were performed, the validations for *ReaderBench* covered a wider spectrum in terms of: 1/ the aggregated cohesion measure by comparison to human evaluations of cohesiveness between adjacent paragraphs; 2/ the scoring mechanism perceived as a summarization facility; 3/ the identification of reading strategies by comparison to the manual scoring scheme; 4/ the textual complexity model emphasizing morphology and semantics factors, compared to the surface metrics used within the DRP score; 5/ participants' involvement in chat conversations with regards to tutor grades; 6/ collaboration assessment through the use of the social knowledge-building model and of the dialogical voice inter-animation model, reflected in the automatic identification of intense collaboration zones; and 7/ aggregation of multiple conversation threads and the validation performed on a long-term discussion group in relation to the critical thinking assessment framework. Moreover, all the previous aspects make the integration of *ReaderBench* appropriate in various educational settings.

In the end, we consider that the initial goal to enhance understanding as a "mediator of learning" by providing feedback to both learners and tutors has been addressed through the multitude of learning tasks that are supported and can be more easily achieved through the use of the developed systems, corroborated with the reliability of the validated outputs.

11.2 Directions for Future Research

As our entire research was inter-disciplinary, the further research directions can be similarly regarded from multiple perspectives. Therefore, from an *educational* point of view, the first thing that needs to be addressed consists of performing the envisioned educational scenarios as controlled experiments with tutors and learners in classroom environments. Secondly, an interesting experiment would envisage a generalization of chat analysis to classroom discussions in order to evaluate interaction in classroom environments between teachers and students. In this context, cohesiveness of the ongoing discussions, voice inter-animation and learner participation/ involvement assessment could become indicators of the ongoing pedagogical activities. Thirdly, a major impact in terms of textual complexity evaluation would play the use of a human-rated corpus as the analysis would become attuned with human perceptions of complexity. At this moment we are currently negotiating with a French publisher for an exchange of corpora and automatic document evaluations in order to enhance our analysis. As a fourth already ongoing research direction, we target an expansion of the long-term discussion groups' evaluation in terms of a more profound personalization of the analysis for addressing virtual communities of practice.

From a *technical* perspective, the first future step will consist of integrating within *ReaderBench* 1/ emotions detection, mood and opinion mining (Lupan, Dascalu, Trausan-Matu, & Dessus, 2012; Lupan, Dascalu, Trausan-Matu, Rebedea, & Dessus, 2012; Lupan, Bobocescu-Kesikis, Dascalu, Trausan-Matu, & Dessus, in press) and 2/ the identification of psychological or personality traits (Ciubuc, Dascalu, Trausan-Matu, & Rebedea, 2012; Ciubuc, Dascalu, Trausan-Matu, & Marhan, 2013) in order to augment the identification of voices through the presentation of a personal point of view. Moreover, valence shifters (Musat & Trausan-Matu, 2010) and argumentation acts are envisioned to be used in order to refine inter-animation detection spanning throughout a conversation. Secondly, the semantic models need to be further improved in terms of disambiguation and outliers cleaning (Musat, Velcin, Rizoiu, & Trausan-Matu, 2011; Musat, Velcin, Trausan-Matu, & Rizoiu, 2011). Thirdly, speech-to-text functionalities would enable the use of *ReaderBench*

with younger pupils and make the software more practicable, although the intrinsic limitations of such automatic transcription services. Nevertheless, punctual improvements will be also taken into consideration: 1/ a deeper filtering of the inferred knowledge concepts with regards to the context; 2/ integration of additional textual complexity factors; or 3/ refinement of the bridging identification strategy.

In addition, from a *philosophical* point of view we aim to extend the current analysis towards exploring the concept of intertextuality (Allen, 2000), tightly related to Bakhtin's dialogism (Bakhtin, 1981), therefore enlarging the perspective from a single discourse, towards the identification of voices that span throughout multiple documents and conversations, addressing the interconnections between them. Moreover, from a completely different point of view, we have already started discussing about the facilities of an additional environment, *WriterBench*, integrating the same core mechanics of *ReaderBench*, but focused on providing just-in-time feedback to learners in order to improve their ongoing learning experience, both while self-explaining and while actively involving themselves in ongoing conversations.

List of Publications

Conferences

Nistor, N., Dascalu, M., Trausan-Matu, S., Mihaila, D., Baltes, B., & Smeaton, G. (in press). Virtual Communities of Practice in Academia: Automated Analysis of Collaboration Based on the Social Knowledge Building Model. In *8th European Conference on Technology Enhanced Learning (EC-TEL 2013)*. Paphos, Cyprus. (ISI Web of Knowledge, Category B in ARC Conference Ranking)

Emin, V., Wasson, B., Hansen, C., Mor, Y., Rodríguez-Triana, M.J., Dascalu, M., . . . Pernin, J.-P. (in press). Towards An Integrated Model of Teacher Inquiry into Student Learning, Learning Design, and Learning Analytics. In *8th European Conference on Technology Enhanced Learning (EC-TEL 2013)*. Paphos, Cyprus. (ISI Web of Knowledge, Category B in ARC Conference Ranking)

Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., & Nardy, A. (2013). ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. In H. C. Lane, K. Yacef, J. Mostow & P. Pavlik (Eds.), *16th Int. Conf. on Artificial Intelligence in Education (AIED 2013)* (pp. 379–388). Memphis, USA: Springer.

Dascalu, M., Trausan-Matu, S., & Dessus, P. (2013). Cohesion-based Analysis of CSCL Conversations: Holistic and Individual Perspectives. In N. Rummel, M. Kapur, M. Nathan & S. Puntambekar (Eds.), *10th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2013)* (pp. 145–152). Madison, USA: ISLS. (ISI Web of Knowledge, Category A in ARC Conference Ranking)

Ciobanu, R.I., Dobre, C., Dascalu, M., Trausan-Matu, S., & Cristea, V. (2013). Collaborative Selfish Node Detection with an Incentive Mechanism for Opportunistic Networks. In *5th IFIP/IEEE International Workshop on Management of the Future Internet in conjunction with IFIP/IEEE International Symposium on Integrated Network Management* (pp. 1161–1166). Ghent, Belgium: IEEE. (ISI Web of Knowledge, Category A in ARC Conference Ranking)

Ciubuc, C., Dascalu, M., Trausan-Matu, S., & Marhan, A.-M. (2013). Forming Teams by Psychological Traits An Effective Method of Developing Groups in an Educational Environment. In I. Dumitrache, A. M. Florea & F. Pop (Eds.), *1st Int. Workshop on Design and Spontaneity in Computer-Supported Collaborative Learning (DS-CSCL-2013), in conjunction with the 19th Int. Conf. on Control Systems and Computer Science (CSCS19)* (pp. 597–604). Bucharest, Romania: IEEE.

Lupan, D., Bobocescu-Kesikis, S., Dascalu, M., Trausan-Matu, S., & Dessus, P. (2013). Predicting Readers' Emotional States Induced by News Articles through Latent Semantic Analysis. In *Social Media in Academia: Research and Teaching (SMART 2013)* (pp. 78–86). Bacau, Romania.

Nistor, N., Baltes, B., Mihaila, D., Dascalu, M., Smeaton, G., & Trausan-Matu, S. (in press). *Virtual communities of practice in academia: An automated analysis of expertise and expert status*. Paper presented at the 15th Biennial EARLI Conference for Research on Learning and Instruction, Munich, Germany.

Dascalu, M., Dessus, P., Bianco, M., Loiseau, M., & Trausan-Matu, S. (2013). *L'apport du TAL dans des environnements favorisant l'apprentissage auto-régulé*. Paper presented at Journée EIAH&IA, Toulouse, France. http://hal.archives-ouvertes.fr/docs/00/82/42/85/PDF/eiahia2013_attachment_5.pdf

Nistor, N., Baltes, B., Smeaton, G., Dascalu, M., Mihaila, D., & Trausan-Matu, S. (2013). *Virtual Communities of Practice in Academia: An Automated Discourse Analysis*. Paper presented at the 1st International Workshop on Discourse-Centric Learning Analytics (DCLA13), a pre-conference workshop at Learning Analytics and Knowledge (LAK2013), Leuven, Belgium. http://www.solaresearch.org/events/lak/lak13/dcla13/

Trausan-Matu, S., Dascalu, M., & Dessus, P. (2012). Textual Complexity and Discourse Structure in Computer-Supported Collaborative Learning. In S. A. Cerri, W. J. Clancey, G. Papadourakis & K. Panourgia (Eds.), *11th Int. Conf. on Intelligent Tutoring Systems (ITS 2012)* (pp. 352–357). Chania, Grece: Springer. (ISI Web of Knowledge, Category A in ARC Conference Ranking)

Trausan-Matu, S., Dascalu, M., & Rebedea, T. (2012). A system for the automatic analysis of Computer-Supported Collaborative Learning chats. In C. Giovannella, D. G. Sampson & I. Aedo (Eds.), *12th IEEE Int. Conf. on Advanced Learning Technologies (ICALT 2012)* (pp. 95–99). Rome, Italy: IEEE. (ISI Web of Knowledge, Category B in ARC Conference Ranking)

Donciu, M., Ionita, M., Dascalu, M., & Trausan-Matu, S. (2012). Ant Colony Optimisation for Automatically Populating Ontologies with Individuals. In *14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*. Timisoara, Romania: IEEE. (ISI Web of Knowledge, Category C in ARC Conference Ranking)

Dagadita, M., Bancu, C., Dascalu, M., Dobre, C., Trausan-Matu, S., & Florea, A.M. (2012). ARSYS - Article Recommender System. In *14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*. Timisoara, Romania: IEEE. (ISI Web of Knowledge, Category C in ARC Conference Ranking)

Dascalu, M., Trausan-Matu, S., & Dessus, P. (2012). Towards an integrated approach for evaluating textual complexity for learning purposes. In E. Popescu, R. Klamma, H. Leung & M. Specht (Eds.), *11th Int. Conf. in Advances in Web-Based Learning (ICWL 2012)* (pp. 268–278). Sinaia, Romania: Springer. (ISI Web of Knowledge)

Lupan, D., Dascalu, M., Trausan-Matu, S., & Dessus, P. (2012). Analyzing emotional states induced by news articles with Latent Semantic Analysis. In A. Ramsay & G. Agre (Eds.), *15th Int. Conf. on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2012)* (pp. 59–68). Varna, Bulgaria: Springer. (ISI Web of Knowledge)

Daniil, C., Dascalu, M., & Trausan-Matu, S. (2012). Automatic forum analysis: a thorough method of assessing the importance of posts, discussion threads and of users' involvement. In D. D. Burdescu, R. Akerkar & C. Badica (Eds.), *2nd Int. Conf. on Web Intelligence, Mining and Semantics (WIMS '12)* (pp. 37). Craiova, Romania: ACM. (ISI Web of Knowledge)

Dessus, P., Bianco, M., Nardy, A., Toffa, F., Dascalu, M., & Trausan-Matu, S. (2012). Automated analysis of pupils' self-explanations of a narrative text. In E. de Vries & K. Scheiter (Eds.), *Staging knowledge and experience, Meeting of the EARLI SIG 2 "Comprehension of Text and Graphics"* (pp. 52–54). Grenoble, France: LSE, Pierre-Mendès-France University.

Dascalu, M., Rebedea, T., Trausan-Matu, S., & Armitt, G. (2011). PolyCAFe: Collaboration and Utterance Assessment for Online CSCL Conversations. In H. Spada, G. Stahl, N. Miyake & N. Law (Eds.), *9th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2011)* (pp. 781–785). Hong Kong, China: ISLS. (ISI Web of Knowledge, Category A in ARC Conference Ranking)

Rebedea, T., Dascalu, M., Trausan-Matu, S., Armitt, G., & Chiru, C.G. (2011). Automatic Assessment of Collaborative Chat Conversations with PolyCAFe. In C. D. Kloos, D. Gillet, R. M. Crespo García, F. Wild & M. Wolpers (Eds.), *Towards Ubiquitous Learning - 6th European Conference of Technology Enhanced Learning (EC-TEL 2011)* (pp. 299–312). Palermo, Italy: Springer. (ISI Web of Knowledge, Category B in ARC Conference Ranking)

Dascalu, M., Dobre, C., Trausan-Matu, S., & Cristea, V. (2011). Beyond Traditional NLP: A Distributed Solution for Optimizing Chat Processing. In *10th International Symposium on Parallel and Distributed Computing (ISPDC 2011)* (pp. 133–138). Cluj-Napoca, Romania: IEEE. (ISI Web of Knowledge, Category C in ARC Conference Ranking)

Donciu, M., Ionita, M., Dascalu, M., & Trausan-Matu, S. (2011). The Runner - Recommender system of workout and nutrition for runners. In D. Wang, V. Negru, T. Ida, T. Jebelean, D. Petcu, S. M. Watt & D. Zaharie (Eds.), *3th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2011)* (pp. 230–238). Timisoara, Romania: IEEE. (ISI Web of Knowledge, Category C in ARC Conference Ranking)

Rebedea, T., Dascalu, M., Trausan-Matu, S., & Chiru, C.G. (2011). Automatic Feedback and Support for Students and Tutors Using CSCL Chat Conversations. In S. Trausan-Matu (Ed.), *First International K-Teams Workshop on Semantic and Collaborative Technologies for the Web* (pp. 20–33). Bucharest, Romania: Politehnica Press.

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. In M. Wolpers, P. Kirschner, M. Scheffel, S. Lindstaedt & V. Dimitrova (Eds.), *Sustaining TEL: From Innovation to Learning and Practice - 5th European Conference on Technology Enhanced Learning (EC-TEL 2010)* (pp. 420–425). Barcelona, Spain: Springer. (ISI Web of Knowledge, Category B in ARC Conference Ranking)

Dascalu, M., Rebedea, T., & Trausan-Matu, S. (2010). A deep insight in chat analysis: Collaboration, evolution and evaluation, summarization and search. In D. Dochev & D. Dicheva (Eds.), *14th Int. Conf. on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2010)* (pp. 191–200). Varna, Bulgaria: Springer. (ISI Web of Knowledge)

Rebedea, T., Dascalu, M., Posea, V., & Trausan-Matu, S. (2010). Web Services for Supporting the Interactions of Learners in the Social Web. In *9th IEEE Int. Conf. RoEduNet* (pp. 128–133). Sibiu, Romania: IEEE. (ISI Web of Knowledge)

Dessus, P., Trausan-Matu, S., Mandin, S., Rebedea, T., Zampa, V., Dascalu, M., & Villiot-Leclercq, E. (2010). *Assessing writing and collaboration in learning: Methodological issues.* Paper presented at the

Workshop "Analysing the quality of collaboration in task-oriented computer-mediated interactions", held in conjunction to the 9th Int. Conf. on the Design of Cooperative Systems (COOP 2010), Aix-en-Provence, France.

Rebedea, T., Dascalu, M., & Trausan-Matu, S. (2010). PolyCAFe: Polyphony-Based System for Collaboration Analysis and Feedback Generation. In S. Trausan-Matu & P. Dessus (Eds.), *Second Workshop on Natural Language in Support of Learning: Metrics, Feedback and Connectivity* (pp. 21–34). Bucharest, Romania: MatrixRom.

Dessus, P., Trausan-Matu, S., Zampa, V., Rebedea, T., Mandin, S., & Dascalu, M. (2009). Vers un environnement-tuteur d'apprentissage dialogique. In C. Develotte, F. Mangenot & E. Nissen (Eds.), *2e Colloque Echanger pour Apprendre en Ligne (EPAL'09)*. Grenoble: LIDILEM/INRP.

Saracin, C.G., Saracin, M., Dascalu, M., & Lepar, A.M. (2008). Spectral analysis of the digital signals - the echo cancellation using the LMS algorithm. In *International Symposium ATEE (Advanced Topics in Electrical Engineering), Section Virtual Instrumentation* (pp. 288–291). Bucharest, Romania: Printech.

Dascalu, M., Chioasca, E.V., & Trausan-Matu, S. (2008). ASAP – An Advanced System for Assessing Chat Participants. In D. Dochev, M. Pistore & P. Traverso (Eds.), *13th Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA 2008)* (pp. 58–68). Varna, Bulgaria: Springer. (ISI Web of Knowledge)

Saracin, C.G., Saracin, M., Dascalu, M., & Caras, V. (2007). *Legatura Matlab-Excel-HPVEE in Generarea de Forme de Unda Arbitrare Utilizate in Studiul Regimului Deformant*. Paper presented at the 4th Int. Conf. METSIM, Bucharest, Romania.

Dascalu, M., Chioasca, E.V., & Trausan-Matu, S. (2007). *Sistem de evaluare a competentei participantilor la un chat*. Paper presented at the Simpozionul Teoria si Practica Invatarii Colaborative in Echipe Virtuale pe Web, Bucharest, Romania.

Journals

Dascalu, M., Trausan-Matu, S., Dessus, P., Bianco, M., & Nardy, A. (in press). ReaderBench, o platformă integrată pentru analiza complexității textuale și a strategiilor de lectură, Revista Romana de Interactiune Om-Calculator. (CNCSIS B+)

Dascalu, M., Trausan-Matu, S., & Dessus, P. (in press). ReaderBench - analiza bazată pe coeziune a implicării și a colaborării participanților în cadrul conversațiilor CSCL. Revista Romana de Interactiune Om-Calculator. (CNCSIS B+)

Ciobanu, R.I., Dobre, C., Dascalu, M., Trausan-Matu, S., & Cristea, V. (in press). SENSE: A Collaborative Selfish Node Detection and Incentive Mechanism for Opportunistic Networks. *Journal of Network and Computer Applications*. (ISI Impact factor 1.065)

Oprescu, B., Dascalu, M., Trausan-Matu, S., Dessus, P., & Bianco, M. (in press). Automated Assessment of Paraphrases in Pupil's Self-Explanations. *Scientific Bulletin, University Politehnica of Bucharest, Series C*. (CNCSIS B+)

Lupan, D., Dascalu, M., Trausan-Matu, S., Rebedea, T., & Dessus, P. (2012). Analiza starilor emotionale induse de citirea unei stiri utilizand Analiza Semantica Latenta. *Revista Romana de Interactiune Om-Calculator, 5*(2), 103–106. (CNCSIS B+)

Oprescu, B., Dascalu, M., Trausan-Matu, S., Rebedea, T., Dessus, P., & Bianco, M. (2012). Analiza automata a auto-explicatiilor. *Revista Romana de Interactiune Om-Calculator, 5*(2), 71–76. (CNCSIS B+)

Ciubuc, C., Dascalu, M., Trausan-Matu, S., & Rebedea, T. (2012). Formarea de echipe pe baza profilelor psihologice. *Revista Romana de Interactiune Om-Calculator, 5*(2), 97–102. (CNCSIS B+)

Rebedea, T., Dascalu, M., Trausan-Matu, S., Teodorescu, A., & Ene, A. (2011). PolyCAFE: Sistem avansat de evaluare a conversatiilor de tip chat bazat pe modelul polifonic. *Revista Romana de Interactiune Om-Calculator, 4*(3), 35–42. (CNCSIS B+)

Dascalu, M., Trausan-Matu, S., Rebedea, T., Lupan, D., & Bobocescu-Kesikis, S. (2011). Concepte specifice dialogismului reliefate în evaluarea colaborării participanților unei discuții de tip chat. *Revista Romana de Interactiune Om-Calculator, 4*(3), 29–34. (CNCSIS B+)

Dascalu, M., Trausan-Matu, S., Rebedea, T., & Daniil, C. (2011). A Dialogic Model for Assessing the Collaboration and the Involvement of Chat Participants. *Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information, 4*(1), 51–58.

Trausan-Matu, S., Dascalu, M., Rebedea, T., & Gartner, A. (2010). Corpus de conversatii multi-participant si editor pentru adnotarea lui. *Revista Romana de Interactiune Om-Calculator, 3*(1), 53–64. (CNCSIS B+)

Dascalu, M., Trausan-Matu, S., & Dessus, P. (2010). Utterances assessment in chat conversations. *Research in Computing Science, 46*, 323–334.

Dascalu, M., & Trausan-Matu, S. (2009). Ch.A.M.P. – A Program for Chat Modelling and Assesment. *Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information, 2*(1), 95–106.

Dascalu, M., & Trausan-Matu, S. (2009). Ch.A.M.P. – Sistem pentru evaluarea si modelarea contributiei participantilor la un Chat. *Revista Romana de Interactiune Om-Calculator, 2*(2), 131–146. (CNCSIS B+)

Saracin, C.G., Saracin, M., Dascalu, M., & Lepar, A.M. (2009). Echo Cancellation Using the LMS Algorithm. *Scientific Bulletin, University Politehnica of Bucharest, Series C, 71*(4), 167–174. (CNCSIS B+)

Dascalu, M., Trausan-Matu, S., & Dessus, P. (2010). Evaluarea si sumarizarea automata a conversatiilor chat. *Revista Romana de Interactiune Om-Calculator, 3*(3), 95–100. (CNCSIS B+)

Dascalu, M., Chioasca, E.V., & Trausan-Matu, S. (2008). ASAP – Sistem avansat de evaluare a participantilor la un chat. *Revista Romana de Interactiune Om-Calculator, 1*(3), 105–112. (CNCSIS B+)

Book Chapters

Trausan-Matu, S., Rebedea, T., & Dascalu, M. (2010). Analysis of discourse in collaborative Learning Chat Conversations with Multiple Participants. In D. Tufis & C. Forascu (Eds.), *Multilinguality and*

Interoperability in Language Processing with Emphasis on Romanian (pp. 313–330). Bucharest, Romania: Editura Academiei.

Project Deliverables

Trausan-Matu, S., Dessus, P., Rebedea, T., Loiseau, M., Dascalu, M., Mihaila, D., . . . Dulceanu, A. (2011). Deliverable D5.3 LTfLL – Learning support and feedback. Heerlen: OUNL.

Trausan-Matu, S., Dessus, P., Rebedea, T., Mandin, S., Villiot-Leclercq, E., Dascalu, M., . . . Graziani, E. (2010). Deliverable D5.2 LTfLL – Learning support and feedback. Heerlen: OUNL.

Invited Talks

Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., & Nardy, A. (2013, scheduled April, 30). ReaderBench – an Environment for Analyzing Textual Complexity, Reading Strategies and Collaboration. Laboratoire des Sciences de l'Education. Grenoble, UPMF, France.

Dascalu, M., Trausan-Matu, S., & Dessus, P. (2013, Jan. 22). ReaderBench - Discourse Analysis Platform. LIG–MeTAH - Modèles et Technologies pour l'Apprentissage Humain. Grenoble, UJF, France.

Dascalu, M., Lupan, D., Trausan-Matu, S., & Dessus, P. (2012, May 31). Analyse automatique des états émotionnels consécutifs à la lecture d'articles de presse. Les 3e Rencontres du Pôle Grenoble Cognition. Grenoble, France.

Trausan-Matu, S., & Dascalu, M. (2011, Nov. 22). PolyCAFe - A system for analyzing participation and collaboration in CSCL online chats [Research seminar]. Laboratoire des Sciences de l'Education. Grenoble, UPMF, France.

References

Abrams, E. (2000). *Topic Sentences and Signposting*. Harvard University. Writing Center. Retrieved from http://www.fas.harvard.edu/~wricntr/documents/TopicSentences.html

Adams, P.H., & Martell, C.H. (2008). Topic Detection and Extraction in Chat. In *IEEE Int. Conf. on Semantic Computing (ICSC 2008)* (pp. 581–588). Santa Clara, CA: IEEE.

Agirre, E., & Lopez, O. (2003). Clustering WordNet Word Senses. In *Conference on Recent Advances on Natural Language (RANLP'03)* (pp. 121–130). Borovetz, Bulgaria: ACL.

Alderson, J. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alias-i. (2008). LingPipe (Version 4.1.0). Retrieved from http://alias-i.com/lingpipe

Allen, G. (2000). *Intertextuality (The New Critical Idiom)*. London, UK: Routledge.

Anderson, J.R. (1985). *Cognitive psychology and its implications*. New York, NY: Freeman.

Arora, R., & Ravindran, B. (2008). Latent dirichlet allocation based multi-document summarization. In *2nd Workshop on Analytics for Noisy Unstructured Text Data* (pp. 91–97). Singapore: ACM.

Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater V.2.0. In *Annual Meeting of the International Association for Educational Assessment*. Philadelphia, PA: Association for Educational Assessment.

Austin, J.L. (1962). *How to Do Things With Words*. Cambridge, MA: Harvard University Press.

Avouris, N., Fiotakis, G., Kahrimanis, G., Margaritis, M., & Komis, V. (2007). Beyond Logging of Fingertip Actions: Analysis of Collaborative Learning Using Multiple Sources of Data. *Journal of Interactive Learning Research, 18*(2), 231–250.

Babu, S. (2010). Towards automatic optimization of MapReduce programs. In *1st ACM Symposium on Cloud Computing (SoCC '10)* (pp. 137–142). Indianapolis, IN: ACM.

Bagozzi, R.P. (2007). The legacy of the Technology Acceptance Model and a proposal for a paradigm shift. *Journal of the Association for Information Systems, 8*(4), 244–254.

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.

Bakhtin, M.M. (1986). *Speech genres and other late essays* (V. W. McGee, Trans.). Austin: University of Texas.

References

Banerjee, S., & Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet. In *Third Int. Conf. on Intelligent Text Processing and Computational Linguistics* (pp. 136–145). Mexico City, Mexico.

Bannister, M.J., Eppstein, D., Goodrich, M.T., & Trott, L. (2012). Force-directed graph drawing using social gravity and scaling. In *20th Int. Symp. Graph Drawing* (pp. 414–425). Redmond, Washington: Springer.

Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In *ACL Workshop on Intelligent Scalable Text Summarization (ISTS'97)* (pp. 10–17). Madrid, Spain: ACL.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media* (pp. 361–362). San Jose, CA: AAAI Press.

Batagelj, V., & Mrvar, A. (2000). Some analyses of Erdős collaboration graph. *Social Networks, 22*(2), 173–186.

Beene, L. (1988). How can functional documents be made more cohesive and coherent? In L. Beene & P. White (Eds.), *Solving problems in technical writing* (pp. 108–129). Oxford: Oxford University Press.

Benjamin, R.G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*, 63–88.

Bereiter, C. (2002). *Education and Mind in the Knowledge Age*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *The Journal of Machine Learning Research, 13*, 281–305.

Berlanga, A.J., Van Rosmalen, P., Trausan-Matu, S., Monachesi, P., & Burek, G. (2009). The Language Technologies for Lifelong Learning Project. In *9th IEEE Int. Conf. on Advanced Learning Technologies (ICALT 2009)* (pp. 624–625). Riga, Latvia: IEEE.

Berliner, P. (1994). *Thinking in jazz: The infinite art of improvisation*. Chicago, IL: University of Chicago Press.

Berry, M.W., Drmac, Z., & Jessup, E.R. (1999). Matrices, vector spaces, and information retrieval. *SIAM Review, 41*, 335–362.

Berry, M.W., Dumais, S.T., & O'Brien, G.W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review, 37*, 573–595.

Bestgen, Y. (2012). Évaluation automatique de textes et cohésion lexicale. *Discours, 11*. doi: 10.4000/discours.8724

Biggs, N., Lloyd, E., & Wilson, R. (1986). *Graph Theory, 1736-1936*. Oxford: Oxford University Press.

Blanc, N., & Brouillet, P. (2003). *Mémoire et compréhension: Lire pour comprendre*. Paris, France: Editions InPress.

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D.M., & Lafferty, J. (2009). Topic Models. In A. Srivastava & M. Sahami (Eds.), *Text Mining: Classification, Clustering, and Applications* (pp. 71–93). London, UK: Chapman & Hall/CRC.

References

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

Boekarts, M., Pintrich, P.R., & Zeidner, M. (Eds.). (2000). *Handbook of self-regulation.* San Diego, CA: Academic Press.

Borgatti, S.P., Mehra, A., Brass, D.J., & Labianca, G. (2009). Network analysis in the social sciences. *Science, 323*(5916), 892–895.

Borman, G.D., & Dowling, N.M. (2004). Testing the Reading Renaissance Program Theory: A Multilevel Analysis of Student and Classroom Effects on Reading Achievement. Madison, WI: University of Wisconsin-Madison.

Bormuth, J.R. (1966). Readability: A new approach. *Reading Research Quarterly, 1*(3), 79–132.

Bormuth, J.R. (1969). Development of readability analysis. Washington, D.C.: U.S. Office of Education, Bureau of Research, U.S. Department of Health, Education, and Welfare.

Borthakur, D., Gray, J., Sarma, J.S., Muthukkaruppan, K., Spiegelberg, N., Kuang, H., . . . Aiyer, A. (2011). Apache Hadoop goes realtime at Facebook. In *ACM SIGMOD Int. Conf. on Management of Data (SIGMOD '11)* (pp. 1071–1080). Athens, Greece: ACM.

Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology, 25*(2), 163–177.

Brown, J.D. (1998). An EFL readability index. *JALT Journal, 20*(2), 7–36.

Bruner, J. (1990). *Acts of meaning.* Cambridge, MA: Harvard University Press.

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources, Second meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 29–34). Pittsburgh, PA.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics, 32*(1), 13–47.

Burstein, J., Kaplan, R., Wolff, S., & Lu, C. (1996). Using lexical semantic techniques to classify free-responses. In *ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons.* Santa Cruz, CA: ACL.

Cassirer, E. (1953). *The philosophy of symbolic forms* (Vol. 1). New Haven, CT: Yale University Press.

Cazden, C.B. (1993). Vygotsky, Hymes, and Bakhtin: From word to utterance and voice. In E. A. Forman, N. Minick & C. A. Stone (Eds.), *Contexts for learning: Sociocultural dynamics in children's development* (pp. 197–212). Oxford: Oxford University Press.

Celce-Murcia, M., & Olshtain, E. (2000). *Discourse and Context in Language Teaching: a Guide for Language Teachers.* New York, NY: Cambridge University Press.

Cha, S.H. (2007). Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences, 1*(4), 300–307.

Chafe, W. (1997). Polyphonic topic development. In T. Givon (Ed.), *Conversation: Cognitive, communicative and social perspectives* (pp. 41–53). Amsterdam, The Netherlands: John Benjamins.

References

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 27:21–27:27.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D.M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams & A. Culotta (Eds.), *23rd Annual Conference on Neural Information Processing Systems (NISP 2009)* (pp. 288–296). Vancouver, Canada.

Chen, M. (1995). A methodology for characterizing computer-based learning environments. *Instructional Science, 23*, 183–220.

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

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater's performance on TOEFL essays. Princeton, NJ: Educational Testing Service.

Chu, C., Kim, S.K., Lin, Y.A., Yu, Y., Bradski, G., Ng, A.Y., & Olukotun, K. (2007). Map-reduce for machine learning on multicore. In B. Scholkopf, J. C. Platt & T. Hofmann (Eds.), *Conference of Advances in Neural Information Processing Systems* (pp. 281–288). Cambridge, MA: MIT Press.

Ciubuc, C., Dascalu, M., Trausan-Matu, S., & Marhan, A.-M. (2013). Forming Teams by Psychological Traits An Effective Method of Developing Groups in an Educational Environment. In *1st International Workshop on Design and Spontaneity in Computer-Supported Collaborative Learning (DS-CSCL-2013), in conjunction with the 19th Int. Conf. on Control Systems and Computer Science (CSCS19)*. Bucharest, Romania: IEEE.

Ciubuc, C., Dascalu, M., Trausan-Matu, S., & Rebedea, T. (2012). Formarea de echipe pe baza profilelor psihologice. *Revista Romana de Interactiune Om-Calculator, 5*(2), 97–102.

Coleman, T.F., & Moré, J.J. (1983). Estimation of sparse Jacobian matrices and graph coloring Problems. *SIAM Journal on Numerical Analysis, 20*(1), 187–209.

Collins, M. (2002). Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In *EMNLP '02 - ACL '02 Conference on Empirical methods in natural language processing* (pp. 1–8). Philadelphia, PA: ACL.

Cooper, A. (2012). A Brief History of Analytics (Vol. 1). Bolton, UK: CETIS Analytics Series.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., & Stein, C. (Eds.). (2009). *Introduction to Algorithms* (3rd ed.). Cambridge, MA: MIT Press.

Cortes, C., & Vapnik, V.N. (1995). Support-Vector Networks. *Machine Learning, 20*(3), 273–297.

Cree, G.S., & Armstrong, B.C. (2012). Computational Models of Semantic Memory. In M. Spivey, K. McRae & M. Joanisse (Eds.), *The Cambridge Handbook of Psycholinguistics* (pp. 259–282). Cambridge: Cambridge University Press.

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.

Dahl, R. (2007). *Matilda* (H. Robillot, Trans.). Paris, France: Gallimard.

Dangalchev, C. (2006). Residual closeness in networks. *Physica A, 365*(2), 556–564.

References

Dascalu, M., Chioasca, E.V., & Trausan-Matu, S. (2008a). ASAP – An Advanced System for Assessing Chat Participants. In D. Dochev, M. Pistore & P. Traverso (Eds.), *13th Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA 2008)* (pp. 58–68). Varna, Bulgaria: Springer.

Dascalu, M., Chioasca, E.V., & Trausan-Matu, S. (2008b). ASAP – Sistem avansat de evaluare a participantilor la un chat. *Revista Romana de Interactiune Om-Calculator, 1*(3), 105–112.

Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., & Nardy, A. (in press). ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. In *16th Int. Conf. on Artificial Intelligence in Education (AIED 2013)*. Memphis, USA: Springer.

Dascalu, M., Dobre, C., Trausan-Matu, S., & Cristea, V. (2011). Beyond Traditional NLP: A Distributed Solution for Optimizing Chat Processing. In *10th International Symposium on Parallel and Distributed Computing (ISPDC 2011)* (pp. 133–138). Cluj-Napoca, Romania: IEEE.

Dascalu, M., Rebedea, T., & Trausan-Matu, S. (2010). A deep insight in chat analysis: Collaboration, evolution and evaluation, summarization and search. In D. Dochev & D. Dicheva (Eds.), *14th Int. Conf. on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2010)* (pp. 191–200). Varna, Bulgaria: Springer.

Dascalu, M., Rebedea, T., Trausan-Matu, S., & Armitt, G. (2011). PolyCAFe: Collaboration and Utterance Assessment for Online CSCL Conversations. In H. Spada, G. Stahl, N. Miyake & N. Law (Eds.), *9th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2011)* (pp. 781–785). Hong Kong, China: ISLS.

Dascalu, M., & Trausan-Matu, S. (2009a). Ch.A.M.P. – A Program for Chat Modelling and Assesment. *Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information, 2*(1), 95–106.

Dascalu, M., & Trausan-Matu, S. (2009b). Ch.A.M.P. – Sistem pentru evaluarea si modelarea contributiei participantilor la un Chat. *Revista Romana de Interactiune Om-Calculator, 2*(2), 131–146.

Dascalu, M., Trausan-Matu, S., & Dessus, P. (2010a). Evaluarea si sumarizarea automata a conversatiilor chat. *Revista Romana de Interactiune Om-Calculator, 3*(3), 95–100.

Dascalu, M., Trausan-Matu, S., & Dessus, P. (2010b). Utterances assessment in chat conversations. *Research in Computing Science, 46*, 323–334.

Dascalu, M., Trausan-Matu, S., & Dessus, P. (2012). Towards an integrated approach for evaluating textual complexity for learning purposes. In E. Popescu, R. Klamma, H. Leung & M. Specht (Eds.), *11th Int. Conf. in Advances in Web-Based Learning (ICWL 2012)* (pp. 268–278). Sinaia, Romania: Springer.

Dascalu, M., Trausan-Matu, S., & Dessus, P. (in press). Cohesion-based Analysis of CSCL Conversations: Holistic and Individual Perspectives. In *10th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2013)*. Madison, USA: ISLS.

Davison, A., & Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly, 17*, 187–209.

de Villiers, P. (1974). Imagery and theme in recall of connected discourse. *Journal of Experimental Psychology, 103*(2), 263–268.

References

Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. In *6th Symposium on Operating System Design and Implementation (OSDI'04)* (pp. 137–149). San Francisco, CA: USENIX Association.

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

Degand, L., & Sanders, T. (2002). The impact of relational markers on expository text comprehension in l1 and l2. *Reading and Writing, 15*(7), 739–757.

Dela Rosa, K., & Eskenazi, M. (2011). Self-Assessment of Motivation: Explicit and Implicit Indicators in L2 Vocabulary Learning. In G. Biswas, S. Bull, J. Kay & A. Mitrovic (Eds.), *15th Int. Conf. on Artificial Intelligence in Education (AIED2011)* (pp. 296–303). Auckland, New Zealand: Springer.

Denhière, G., Lemaire, B., Bellissens, C., & Jhean-Larose, S. (2007). A semantic space for modeling children's semantic memory. In T. K. Landauer, D. S. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 143–165). Mahwah: Erlbaum.

Derry, S.J., Hmelo-Silver, C.E., Nagarajan, A., Chernobilsky, E., Feltovich, J., & Halfpap, B. (2005). Making a mesh of it: a STELLAR approach to teacher professional development. In T. Koschmann, D. Suthers & T. W. Chan (Eds.), *Conf. on Computer Supported Collaborative Learning 2005: The Next 10 Years! (CSCL'05)*. Taipei, Taiwan: ISLS.

Dessus, P., Bianco, M., Nardy, A., Toffa, F., Dascalu, M., & Trausan-Matu, S. (2012). Automated analysis of pupils' self-explanations of a narrative text. In E. de Vries & K. Scheiter (Eds.), *Staging knowledge and experience, Meeting of the EARLI SIG 2 "Comprehension of Text and Graphics"* (pp. 52–54). Grenoble, France: LSE, Pierre-Mendès-France University.

Dessus, P., & Trausan-Matu, S. (2010). Implementing Bakhtin's dialogism theory with NLP techniques in distance learning environments. In S. Trausan-Matu & P. Dessus (Eds.), *Proc. 2nd Workshop on Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity (NLPsL 2010)* (pp. 11–20). Bucharest, Romania: Matrix Rom.

Donaway, R.L., Drummey, K.W., & Mather, L.A. (2000). A comparison of rankings produced by summarization evaluation measures. In *Workshop on Automatic summarization (NAACL-ANLP-AutoSum '00)* (pp. 69–78). Stroudsburg, PA: ACL.

Dong, A. (2005). The latent semantic approach to studying design team communication. *Design Studies, 26*(5), 445–461.

Dong, A. (2006). Concept formation as knowledge accumulation: A computational linguistics study. *AIE EDAM: Artificial Intelligence for Engineering Design, Analysis, and Manufacturing, 20*(1), 35–53.

Dong, A. (2009). *The language of design: Theory and computation*. New York, NY: Springer.

Dowell, J., & Gladisch, T. (2007). Design of argument diagramming for case-based group learning. *ACM Int. Conf. Proceeding Series, 250*, 99–105.

References

Dowell, J., Tscholl, M., Gladisch, T., & Asgari-Targhi, M. (2009). Argumentation scheme and shared online diagramming in case-based collaborative learning. In *9th Int. Conf. on Computer supported collaborative learning (CSCL'09)* (pp. 567–575). Rhodes, Greece: ISLS.

Duan, K.-B., & Keerthi, S.S. (2005). Which Is the Best Multiclass SVM Method? An Empirical Study. In *6th International Workshop on Multiple Classifier Systems* (pp. 278–285). Seaside, CA: Springer.

Dumais, S.T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers, 23*(2), 229–236.

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

Dyke, G., Lund, K., & Girardot, J.-J. (2009). Tatiana: an environment to support the CSCL analysis process. In *9th Int. Conf. on Computer supported collaborative learning (CSCL'09)* (pp. 58–67). Rhodes, Greece: ISLS.

Eastman, J.K., & Swift, C.O. (2002). Enhancing collaborative learning: discussion boards and chat rooms as project communication tools. *Business Communication Quarterly, 65*(3), 29–41.

Emin, V., Pernin, J.-P., & Guéraud, V. (2009). Model and tool to clarify intentions and strategies in learning scenarios design. In U. Cress, V. Dimitrova & M. Specht (Eds.), *4th European Conference on Technology Enhanced Learning* (pp. 462–476). Nice, France: Springer.

Emin, V., Pernin, J.-P., Prieur, M., & Sanchez, E. . (2007). Stratégies d'élaboration, de partage et de réutilisation de scénarios pédagogiques. *International Journal of Technologies in Higher Education, 4*(2), 25–37.

Emin, V., Wasson, B., Hansen, C., Mor, Y., Rodríguez-Triana, M.J., Dascalu, M., . . . Pernin, J.-P. (submitted). Towards An Integrated Model of Teacher Inquiry into Student Learning, Learning Design, and Learning Analytics. In *8th European Conference on Technology Enhanced Learning (EC-TEL 2013)*. Paphos, Cyprus.

Emond, B. (2006). WN-LEXICAL: An ACT-R module built from the WordNet lexical database. In *11th Int. Conf. on Cognitive Modeling* (pp. 359–360). Trieste, Italy.

Enkvist, N.E. (1987). Text Linguistics for the Applier: An Orientation. In U. Connor & R. B. Kaplan (Eds.), *Writing across Languages: Analysis of L2 Text* (pp. 23–44). Reading, MA: Addison Wesley.

Enkvist, N.E. (1990). Seven Problems in the Study of Coherence and Interpretability. In U. Connor & A. M. Johns (Eds.), *Coherence in writing: Research and Pedagogical Perspectives* (pp. 11–28). Alexandria, VA: TESOL.

Eshelman, L.J. (1991). The CHC Adaptive Search Algorithm: How to Have Safe Search When Engaging in Nontraditional Genetic Recombination. In G. Rawlins (Ed.), *First Workshop on Foundations of Genetic Algorithms* (pp. 265–283). Bloomington, IN: Morgan Kaufmann.

Fano, R.M. (1961). *Transmission of Information: A Statistical Theory of Communication*. Cambridge, MA: MIT Press.

Fellbaum, C. (2005). WordNet(s). In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd ed., Vol. 13, pp. 665–670). Oxford: Elsevier.

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

References

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A Comparison of Features for Automatic Readability Assessment. In *23rd Int. Conf. on Computational Linguistics (COLING 2010)* (pp. 276–284). Beijing, China: ACL.

Finkel, J.R., Grenager, T., & Manning, C.D. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)* (pp. 363–370). Ann Arbor, MI: ACL.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233.

Foltz, P.W., Kintsch, W., & Landauer, T.K. (1993). *An analysis of textual coherence using latent semantic indexing*. Paper presented at the 3rd Annual Conference of the Society for Text and Discourse, Boulder, CO.

Foltz, P.W., Kintsch, W., & Landauer, T.K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25*(2-3), 285–307.

François, F. (1993). *Pratiques de l'oral. Dialogique, jeu et variations de figures du sens*. Paris: Nathan Pédagogie.

François, T. (2012). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère.* (Doctoral dissertation), Université Catholique de Louvain, Faculté de Philosophie, Arts et Lettres, Louvain-la-Neuve, Belgium.

François, T., & Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? In *First Workshop on Predicting and improving text readability for target reader populations (PITR2012)* (pp. 49–57). Montreal, Canada: ACL.

Freeman, L. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry, 40*(1), 35–41.

Freeman, L. (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver: Empirical Press.

Galley, M., & McKeown, K. (2003). Improving Word Sense Disambiguation in Lexical Chaining. In G. Gottlob & T. Walsh (Eds.), *18th International Joint Conference on Artificial Intelligence (IJCAI'03)* (pp. 1486–1488). Acapulco, Mexico: Morgan Kaufmann Publishers, Inc.

Gamallo, P., & Bordag, S. (2011). Is singular value decomposition useful for word similarity extraction? *Language Resources and Evaluation, 45*(2), 95–119.

Garg, R.P., & Sharapov, I. (2002). *Techniques for Optimizing Applications - High Performance Computing*. Upper Saddle River, NJ: Prentice-Hall.

Garrison, D.R., Anderson, T., & Archer, W. (2000). Critical inquiry in a text-based environment: Computer conferencing in higher education. *Internet and Higher Education, 2*(2-3), 87–105.

Garrison, D.R., Anderson, T., & Archer, W. (2010). The first decade of the community of inquiry framework: A retrospective. *Internet and Higher Education, 13*(1-2), 5–9.

Gates, A.F, Natkovich, O., Chopra, S., Kamath, P., Narayanamurthy, S.M., Olston, C., . . . Srivastava, U. (2009). Building a high-level dataflow system on top of Map-Reduce: the Pig experience. *Proceedings of the VLDB Endowment, 2*(2), 1414–1425.

Geisser, S. (1993). *Predictive inference: an introduction*. New York, NY: Chapman and Hall.

References

Genesereth, M.R., & Nilsson, N.J. (1987). *Logical Foundations of Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann Publishers.

Gervasi, V., & Ambriola, V. (2002). Quantitative assessment of textual complexity. In M. L. Barbaresi (Ed.), *Complexity in language and text* (pp. 197–228). Pisa, Italy: Plus.

Glahn, C., Specht, M., & Koper, R. (2008). *Reflecting on web-readings with tag clouds*. Paper presented at the Computer-based Knowledge & Skill Assessment and Feedback in Learning Settings (CAF), special track at the 11th Int. Conf. on Interactive Computer aided Learning (ICL 2008), Villach, Austria.

Glahn, C., Specht, M., & Koper, R. (2009). Reflection support using multi-encoded Tag-clouds. In F. Wild, M. Kalz, M. Palmér & D. Müller (Eds.), *2nd Workshop Mash-Up Personal Learning Envrionments (MUPPLE'09) in conjunction with the 4th European Conference on Technology Enhanced Learning (EC-TEL 2009): Synergy of Disciplines*. Nice, France: CEUR workshop proceedings.

Golub, G.H., & Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis, 2*(2), 205–224.

Golub, G.H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik, 14*(5), 403–420.

Gordon, P.C., Grosz, B.J., & Gillom, L.A. (1993). Pronouns, Names and the Centering of Attention in discourse. *Cognitive Science, 17*(3), 311–347.

Graesser, A.C. (2007). An introduction to strategic reading comprehension. In D. S. McNamara (Ed.), *Reading comprehension strategies: theories, intervention and technologies* (pp. 3–26). Mahwah: Erlbaum.

Graesser, A.C., McNamara, D.S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers, 36*(2), 193–202.

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.

Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371–395.

Green, S., de Marneffe, M.-C., Bauer, J., & Manning, C.D. (2010). Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. In *Conference on Empirical Methods on Natural Language Processing (EMNLP 2011)* (pp. 725–735). Edinburgh, UK: ACL.

Griffiths, T. (2002). Gibbs sampling in the generative model of Latent Dirichlet Allocation. Stanford, CA: Stanford University.

Grosz, B.J., & Sidner, C.L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics, 12*(3), 175–204.

Grosz, B.J., Weinstein, S., & Joshi, A.K. (1995). Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics, 21*(2), 203–225.

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

References

Gunning, R. (1952). *The technique of clear writing*. New York, NY: McGraw-Hill.

Halliday, M.A.K., & Hasan, R. (1976). *Cohesion In English*. London: Longman.

Harary, F. (1979). *Topics in Graph Theory*. New York, NY: New York Academy of Sciences, .

Harrer, A., Hever, R., & Ziebarth, S. (2007). Empowering researchers to detect interaction patterns in e-collaboration. In R. Luckin, K. R. Koedinger & J. E. Greer (Eds.), *13th Int. Conf. on Artificial Intelligence in Education (AIED 2007)* (pp. 503–510). Los Angeles, California, USA: Frontiers in Artificial Intelligence and Applications.

Hasan, R. (1984). Coherence and Cohesive Harmony. In J. Flood (Ed.), *Understanding Reading Comprehension* (pp. 181–219). Newark, DL: International Reading Association.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd Ed.). New York, NY: Springer.

Heer, J., Card, S.K., & Landay, J.A. (2005). Prefuse: a toolkit for interactive information visualization. In G. van der Veer & C. Gale (Eds.), *SIGCHI 2005 Conference on Human factors in Computing Systems (CHI'05)* (pp. 421–430). Portland, OR: ACM.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2006). *Classroom success of an intelligent tutoring system for lexical practice and reading comprehension.* Paper presented at the 9th Int. Conf. on Spoken Language Processing, Pittsburgh, PA.

Heinrich, G. (2008). Parameter estimation for text analysis. Leipzig, Germany: vsonix GmbH + University of Leipzig.

Hirst, G., & St-Onge, D. (1997). Lexical Chains as representation of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hmelo-Silver, C.E., Chernobilsky, E., & Masto, O. (2006). Representations for Analyzing Tool-mediated Collaborative Learning. In *7th Int. Conf. of the Learning Sciences (ICLS '06)* (pp. 1061–1062). Bloomington, IN, USA.

Hmelo-Silver, C.E., Chernobilsky, E., & Nagarajan, A. (2005). *Two sides of the coin: Multiple perspectives on collaborative knowledge construction in online problem-based learning.* Paper presented at the European Association for Research on Learning and Instruction, Nicosia, Cyprus.

Hobbs, J.R. (1978). Why is Discourse Coherent? Menlo Park, California: SRI International.

Hobbs, J.R. (1979). Coherence and Coreference. *Cognitive Science, 3*(1), 67–90.

Hobbs, J.R. (1985). On the Coherence and Structure of Discourse. Center for the Study of Language and Information: Stanford University.

Hobbs, J.R. (1990). Topic drift. In B. Dorval (Ed.), *Conversational Organization and its Development* (pp. 3–22). Norwood, NJ: Ablex Publishing Corp.

Holmer, T., Kienle, A., & Wessner, M. (2006). Explicit Referencing in Learning Chats: Needs and Acceptance. In W. Nejdl & K. Tochtermann (Eds.), *Innovative Approaches for Learning and Knowledge Sharing, First European Conference on Technology Enhanced Learning, EC-TEL 2006* (pp. 170– 184). Crete, Greece: Springer.

References

House, A., & Kuiken, F. (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics, 30*(4), 461 – 473.

Howard, M.W., & Kahana, M.J. (1999). Temporal Associations and Prior-List Intrusions in Free Recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 923–941.

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks, 13*(2), 415–425.

Hsu, C.W., Chang, C.-C., & Lin, C.-J. (2010). A practical guide to support vector classification. Taipei, Taiwan: National Taiwan University.

Hudelot, C. (1994). La circulation interactive du sens dans le dialogue. In A. Trognon, U. Dausendschön-Gay, U. Krafft & C. Riboni (Eds.), *La construction interactive du quotidien*. Nancy, France: Presses Universitaires de Nancy.

Hurme, T.-R., Palonen, T., & Järvelä, S. (2006). Metacognition in joint discussions: An analysis of the patterns of interaction and the metacognitive content of the networked discussions. *Metacognition and Learning, 1*, 181–200.

Jackson, G.T., Guess, R.H., & McNamara, D.S. (2009). Assessing cognitively complex strategy use in an untrained domain. In N. A. Taatgen, H. v. Rijn, L. Schomaker & J. Nerbonne (Eds.), *31st Annual Meeting of the Cognitive Science Society (CogSci '09)* (pp. 2164–2169). Amsterdam, The Netherlands: Cognitive Science Society.

Jadelot, C., Mangeot, M., Petitjean, E., & Salmon-Alt, S. (2006). *Morphalou 2*. Retrieved from: http://www.cnrtl.fr/lexiques/morphalou

Järvelä, S., Hurme, T.-R., & Järvenoja, H. (2011). Self–regulation and motivation in CSCL environments. In S. Ludvigsen, A. Lund & R. Säljö (Eds.), *Learning in social practices: ICT and new artifacts–transformation of social and cultural practices*. Oxford: Pergamon Press.

Järvenoja, H., & Järvelä, S. (2009). Emotion control in collaborative learning situations – Do students regulate emotions evoked from social challenges? *British Journal of Educational Psychology, 79*(3), 463–481.

Jessup, E.R., & Martin, J.H. (2001). Taking a new look at the latent semantic analysis approach to information retrieval. In M. W. Berry (Ed.), *Computational information retrieval* (pp. 121–144). Philadelphia, PA: SIAM.

Jiang, J.J., & Conrath, D.W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Int. Conf. on Research in Computational Linguistics (ROCLING X)* (pp. 19–33). Taipei, Taiwan: Academia Sinica.

Joshi, M., & Rosé, C.P. (2007). Using Transactivity in Conversation Summarization in Educational Dialog. In *SLaTE Workshop on Speech and Language Technology in Education*. Farmington, Pennsylvania, USA.

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

References

Kali, Y., & Linn, M.C. (2007). Technology-enhanced support strategies for inquiry learning. In J. M. Spector, M. D. Merrill, J. J. G. V. Merriënboer & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 445–490). Mahwah, NJ: Erlbaum.

Kaufman, P. (2005). *Smarter Trading: Improving Performance in Changing Markets* (1st Ed.). New York, NY: McGraw-Hill.

Kent, J.T. (1983). Information gain and a general measure of correlation. *Biometrika, 70*(1), 163–173.

Kienle, A., & Wessner, M. (2006). Analyzing and cultivating scientific communities of practice. *International Journal of Web Based Communities, 2*(4), 377–393.

Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975). *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*: Chief of Naval Technical Training, Naval Air Station Memphis.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review, 7*, 257–266.

Kintsch, W. (2001). Predication. *Cognitive Science, 25*(2), 173–202.

Kintsch, W., Kozminsky, E., Streby, W., Mckoon, G., & Keenan, J. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior, 14*(2), 196–214.

Kireyev, K., & Landauer, T.K. (2011). Word Maturity: Computational Modeling of Word Knowledge. In *49th Annual Meeting of the Association for Computational Linguistics* (pp. 299–308). Portland, Oregon: ACL.

Klein, D., & Manning, C.D. (2003). Accurate Unlexicalized Parsing. In *41st Meeting of the Association for Computational Linguistics* (pp. 423–430). Sapporo, Japan: ACL.

Kontostathis, A., Edwards, L., Bayzick, J., McGhee, I., Leatherman, A., & Moore, K. (2009). Comparison of Rule-based to Human Analysis of Chat Logs. In P. Meseguer, L. Mandow & R. M. Gasca (Eds.), *1st International Workshop on Mining Social Media Programme, Conferencia de la Asociación Española Para La Inteligencia Artificial*. Seville, Spain: Springer.

Koschmann, T. (1999). Toward a dialogic theory of learning: Bakhtin's contribution to understanding learning in settings of collaboration. In C. M. Hoadley & J. Roschelle (Eds.), *Int. Conf. on Computer Support for Collaborative Learning (CSCL'99)* (pp. 308–313). Palo Alto: ISLS.

Koslin, B.L., Zeno, S.M., Koslin, S., Wainer, H., & Ivens, S.H. (1987). *The DRP: An effectiveness measure in reading*. New York, NY: College Entrance Examination Board.

Kotz, S., Balakrishnan, N., & Johnson, N.L. (2000). Dirichlet and Inverted Dirichlet Distributions *Continuous Multivariate Distributions* (Vol. 1: Models and Applications). New York, NY: Wiley.

Kozak, K., Agrawal, A., Machuy, N., & Csucs, G. (2009). Data Mining Techniques in High Content Screening: A Survey. *Journal of Computer Science & Systems Biology, 2*, 219–239.

Krzanowski, W.J. (2000). *Principles of Multivariate Analysis: A User's Perspective*. Oxford: Oxford University Press.

Kukemelk, H., & Mikk, J. (1993). The Prognosticating Effectivity of Learning a Text in Physics. *Glottometrica, 14*, 82–103.

References

Kullback, S., & Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics, 22*(1), 79–86.

Kumar, R., Chaudhuri, S., Howley, I., & Rosé, C.P. (2009). VMT-Basilica: an environment for rapid prototyping of collaborative learning environments with dynamic support. In *9th Int. Conf. on Computer supported collaborative learning (CSCL'09)* (pp. 192–194). Rhodes, Greece: ISLS.

Lajoie, S.P., & Azevedo, R. (2006). Teaching and learning in technology-rich environments. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 803–821). Mahwah, NJ: Erlbaum.

Landauer, T.K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *The Psychology of Learning and Motivation, 41*, 43–84.

Landauer, T.K., & Dumais, S. (2008). Latent semantic analysis. *Scholarpedia, 3*(11), 4356.

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.

Landauer, T.K., Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15*(1), 92–108.

Landauer, T.K., Laham, D., & Foltz, P.W. (1998). Learning human-like knowledge by singular value decomposition: a progress report. In M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems, vol. 10* (pp. 45–51). Cambridge, MA: MIT Press.

Landauer, T.K., McNamara, D.S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Lapata, M., & Barzilay, R. (2005). Automatic evaluation of text coherence: models and representations. In *19th international joint conference on Artificial intelligence* (pp. 1085–1090). Edinburgh, Scotland: Morgan Kaufmann Publishers Inc.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Law, N., Lu, J., Leng, J., Yuen, J., & Lai, M. (2008). Understanding Knowledge Building from Multiple Perspectives. In *Workshop on A Common Framework for CSCL Interaction Analysis, ICLS 2008*. Utrecht, Netherland: ISLS.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for wordsense identification. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 265–283). Cambridge, MA: MIT Press.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics, 39*(4).

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *CONLL Shared Task*

'11 Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task (pp. 28–34). Portland, OR: ACL.

Lefebvre, H. (2004). *Rhythmanalysis: Space, Time and Everyday Life* (S. Elden & G. Moore, Trans.). London, UK: Continuum.

Lemaire, B. (2009). Limites de la lemmatisation pour l'extraction de significations. In *9es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2009)*. Lyon, France.

Lennon, C., & Burdick, H. (2004). The Lexile Framework as an approach for reading measurement and success. Durham, NC: MetaMetrics, Inc.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In V. DeBuys (Ed.), *5th Annual Int. Conf. on Systems documentation (SIGDOC '86)* (pp. 24–26). Toronto, Ontario, Canada: ACM.

Lin, D. (1998). An information-theoretic definition of similarity. In *15th Int. Conf. on Machine Learning* (pp. 296–304). Madison, WI, USA: Morgan Kaufmann.

Linell, P. (2001). A dialogical conception of focus groups and social representations. In U. S. Larsoon (Ed.), *Socio-cultural theory and methods: An anthology*. Trollhättan, Sweeden: University of Trollhättan/Uddevalla.

Linell, P. (2005). *The written language bias in linguistics: Its nature, origin and transformations*. Oxford: Routledge.

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

Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* (2nd Ed.). USA: Springer.

Lizza, M., & Sartoretto, F. (2001). A comparative analysis of LSI strategies. In M. W. Berry (Ed.), *Computational information retrieval* (pp. 171–181). Philadelphia, PA: SIAM.

London, J., & Jones, K. (2011). Rhythmic Refinements to the nPVI Measure: A Reanalysis of Patel & Daniele (2003a). *Music Perception: An Interdisciplinary Journal, 29*(1), 115–120.

Low, Y., Bickson, D., Gonzalez, J., Guestrin, C., Kyrola, A., & Hellerstein, J.M. (2012). Distributed GraphLab: a framework for machine learning and data mining in the cloud. *Proceedings of the VLDB Endowment, 5*(8), 716–727.

Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., & Hellerstein, J.M. (2010). GraphLab: A New Parallel Framework for Machine Learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)* (pp. 340–349). Catalina Island, California.

Lupan, D., Bobocescu-Kesikis, S., Dascalu, M., Trausan-Matu, S., & Dessus, P. (in press). Predicting Readers' Emotional States Induced by News Articles through Latent Semantic Analysis. In *Social Media in Academia: Research and Teaching (SMART 2013)*. Bacau, Romania.

Lupan, D., Dascalu, M., Trausan-Matu, S., & Dessus, P. (2012). Analyzing emotional states induced by news articles with Latent Semantic Analysis. In A. Ramsay & G. Agre (Eds.), *15th Int. Conf. on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2012)* (pp. 59–68). Varna, Bulgaria: Springer.

References

Lupan, D., Dascalu, M., Trausan-Matu, S., Rebedea, T., & Dessus, P. (2012). Analiza starilor emotionale induse de citirea unei stiri utilizand Analiza Semantica Latenta. *Revista Romana de Interactiune Om-Calculator, 5*(2), 103–106.

MacKay, D.J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Mahnkopf, C.S. (2002). Theory of Polyphony. In C. S. Mahnkopf, F. Cox & W. Schurig (Eds.), *Polyphony and Complexity*. Hofheim, Germany: Wolke Verlags Gmbh.

Mandin, S. (2009). *Modèles cognitifs computationnels de l'activité de résumer : expérimentation d'un eiah auprès d'élèves de lycée*. (Doctoral dissertation), Université Grenoble-2 - Pierre-Mendès-France, Grenoble, France.

Mann, W.C., & Thompson, S.A. (1987a). Rhetorical structure theory: a theory of text organization. In L. Polanyi (Ed.), *The structure of discourse*. Norwood: Ablex.

Mann, W.C., & Thompson, S.A. (1987b). Rhetorical Structure Theory: A Theory of Text Organization. Marina del Rey, CA: Information Sciences Institute.

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

Manning, C.D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval* (Vol. 1). Cambridge: Cambridge University Press.

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

Marková, I., Linell, P., Grossen, M., & Salazar Orvig, A. (2007). *Dialogue in focus groups: Exploring socially shared knowledge*. London, UK: Equinox.

McCallum, A.K. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved from http://mallet.cs.umass.edu/

McCandless, M., Hatcher, E., & Gospodnetic, O. (2010). *Lucene in Action, Second Edition: Covers Apache Lucene 3.0* (2nd ed.). Greenwich, USA: Manning Publications Co.

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

McNamara, D.S. (Ed.). (2007). *Reading Comprehension Strategies: Theories, Interventions and Technologies*. New York, NY: Erlbaum.

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., Graesser, A.C., & Louwerse, M.M. (in press). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini & E. Albro (Eds.), *Assessing reading in the 21st century*. Lanham, MD: R&L Education.

McNamara, D.S., Kintsch, E., Butler Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1–43.

References

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.

McNamara, D.S., & Magliano, J.P. (2009). Self-explanation and metacognition. In J. D. Hacher, J. Dunlosky & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 60–81). Mahwah, NJ: Erlbaum.

McNamara, D.S., O'Reilly, T.P., Rowe, M., Boonthum, C., & Levinstein, I.B. (2007). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 397–420). Mahwah, NJ: Erlbaum.

McNamara, D.S., & Scott, J. L. (1999). Training reading strategies *21th Annual Meeting of the Cognitive Science Society (CogSci '99)* (pp. 387–392). Hillsdale: Erlbaum.

Medina, R., & Suthers, D. (2008). Bringing Representational Practice From Log to Light. In P. A. Kirschner, F. Prins, V. Jonker & G. Kanselaar (Eds.), *8th Int. Conf. for the Learning Sciences (ICLS 2008)* (pp. 59–66). Utrecht, Netherlands: ISLS.

Metcalfe, J., & Shimamura, A.P. (1994). *Metacognition: knowing about knowing*. Cambridge, MA: MIT Press.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)* (pp. 404–411). Barcelona, Spain: ACL.

Miller, G.A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review, 63*(2), 81–97.

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38*(11), 39–41.

Miller, G.A. (1998). Foreword. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Miller, G.A. (2010). WordNet. Princeton, New Jersey, USA: Princeton University. Retrieved from http://wordnet.princeton.edu

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: R & L Publishing.

Millis, K., Magliano, J.P., & Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and Latent Semantic Analysis. *Scientific Studies of Reading, 10*(3), 225–240.

Milone, M. (2012). The Development of ATOS: The Renaissance Readability Formula. Wisconsin Rapids, WI: Renaissance Learning, Inc.

Miltsakaki, E., & Kukich, K. (2000). The role of centering theory's rough-shift in the teaching and evaluation of writing skills. In *38th Annual Meeting on Association for Computational Linguistics* (pp. 408–415). Hong Kong: ACL.

Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *7th ACM SIGCOMM conference on Internet measurement* (pp. 29–42). San Diego, CA: ACM.

Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press.

Mitchell, T. (1997). *Machine Learning*. New York, NY: McGraw Hill.

References

Moldovan, D.I., & Mihalcea, R. (2000). Using WordNet and Lexical Operators to Improve Internet Searches. *IEEE Internet Computing, 4*(1), 34–43.

Monson, I. (1996). *Saying something: Jazz improvisation and interaction*. Chicago, IL: University of Chicago Press.

Moody, J., & White, D. (2003). Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups. *American Sociological Review, 68*(1), 1–25.

Moore, J.D., & Pollack, M.E. (1992). A problem for RST: the need for multi-level discourse analysis. *Computational Linguistics, 18*(4), 537–544.

Mor, Y., & Craft, B. (2012). Learning design: reflections on a snapshot of the current landscape. *Research in Learning Technology, 20*.

Morris, J., & Hirst, G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics, 17*(1), 21–48.

Muhlenbein, H., & Schlierkamp-Voosen, D. (1993). Predictive models for the breeder genetic algorithm i. continuous parameter optimization i. continuous parameter optimization. *Evolutionary Computation, 1*(1), 25–49.

Murphy, M.L. (2003). *Semantic Relations and the Lexicon: antonymy, synonymy and other paradigms*. Cambridge: Cambridge University Press.

Musat, C.C., & Trausan-Matu, S. (2010). The Impact of Valence Shifters on Mining Implicit Economic Opinions. In D. Dicheva & D. Dochev (Eds.), *Artificial Intelligence: Methodology, Systems, and Applications, 14th Int. Conf. (AIMSA 2010)* (pp. 131–140). Varna, Bulgaria: Springer.

Musat, C.C., Velcin, J., Rizoiu, M.-A., & Trausan-Matu, S. (2011). Concept-Based Topic Model Improvement. In D. Ryzko, H. Rybinski, P. Gawrysiak & M. Kryszkiewicz (Eds.), *Emerging Intelligent Technologies in Industry, 19th International Symposium (ISMIS 2011)* (pp. 133–142). Warsaw, Poland: Springer.

Musat, C.C., Velcin, J., Trausan-Matu, S., & Rizoiu, M.-A. (2011). Improving Topic Evaluation Using Conceptual Knowledge. In T. Walsh (Ed.), *22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)* (pp. 1866–1871). Barcelona, Spain: IJCAI/AAAI.

Nardy, A., Bianco, M., Toffa, F., Rémond, M., & Dessus, P. (in press). Contrôle et régulation de la compréhension: l'acquisition de stratégies de 8 à 11 ans. In J. David & C. Royer (Eds.), *L'apprentissage de la lecture : convergences, innovations, perspectives*. Bern-Paris: Peter Lang.

Nash-Ditzel, S. (2010). Metacognitive Reading Strategies Can Improve Self-Regulation. *Journal of College Reading and Learning, 40*(2), 45–63.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). Common Core State Standards. Washington D.C.: National Governors Association Center for Best Practices, Council of Chief State School Officers.

Navigli, R. (2006). Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *44th Annual Meeting of the Association for Computational Linguistics joint with the 21st Int. Conf. on Computational Linguistics (COLING-ACL 2006)* (pp. 105–112). Sydney, Australia: ACL.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys, 41*(2), 1–69.

References

Navigli, R., & Velardi, P. (2003). An Analysis of Ontology-based Query Expansion Strategies. In *Workshop on Adaptive Text Extraction and Mining (ATEM 2003), in the 14th European Conference on Machine Learning (ECML 2003)* (pp. 42–49). Cavtat-Dubrovnik, Croatia: Springer.

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.

The New Harvard Dictionary of Music. (1986). Cambridge, MA: Harvard University Press.

Newman, M.E.J. (2005). A measure of betweenness centrality based on random walks. *Social Networks, 27*, 39–54.

Newman, M.E.J. (2008). Mathematics of networks. In L. E. Blume & S. N. Durlauf (Eds.), *The New Palgrave Encyclopedia of Economics* (2nd ed.). Basingstoke: Palgrave Macmillan.

Newman, M.E.J. (2010). *Networks: An Introduction* (1st Ed.). Oxford: Oxford University Press.

Nguyen, Quan H., & Hong, Seok-Hee. (2006). Comparison of centrality-based planarisation for 2.5D graph drawing. Sidney: NICTA technical report.

Nistor, N. (2010). Knowledge Communities in the Classroom of the Future. In K. Mäkitalo-Siegl, J. Zottmann, F. Kaplan & F. Fischer (Eds.), *Classroom of the Future - Orchestrating Collaborative Spaces*. Rotterdam: Sense Publishers.

Nistor, N., Baltes, B., Mihaila, D., Dascalu, M., Smeaton, G., & Trausan-Matu, S. (in press). *Virtual communities of practice in academia: An automated analysis of expertise and expert status*. Paper presented at the 15th Biennial EARLI Conference for Research on Learning and Instruction, Munich, Germany.

Nistor, N., Baltes, B., Smeaton, G., Dascalu, M., Mihaila, D., & Trausan-Matu, S. (2013). *Virtual Communities of Practice in Academia: An Automated Discourse Analysis*. Paper presented at the 1st International Workshop on Discourse-Centric Learning Analytics (DCLA13), a pre-conference workshop at Learning Analytics and Knowledge (LAK2013), Leuven, Belgium. http://www.solaresearch.org/events/lak/lak13/dcla13/

Nistor, N., Baltes, B., Smeaton, G., Dascalu, M., Mihaila, D., & Trausan-Matu, S. (submitted). Virtual Communities of Practice in Academia: Participation under the Influence of Technology Acceptance and Community Roles. In *8th European Conference on Technology Enhanced Learning (EC-TEL 2013)*. Paphos, Cyprus.

Nistor, N., Dascalu, M., Trausan-Matu, S., Mihaila, D., Baltes, B., & Smeaton, G. (submitted). Virtual Communities of Practice in Academia: Automated Analysis of Collaboration Based on the Social Knowledge Building Model. In *8th European Conference on Technology Enhanced Learning (EC-TEL 2013)*. Paphos, Cyprus.

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

Nistor, N., Lerche, T., Weinberger, A., Ceobanu, C., & Heymann, O. (2012). Towards the integration of culture into the Unified Theory of Acceptance and Use of Technology. *British Journal of Educational Technology*. doi: 10.1111/j.1467-8535.2012.01383.x

References

O'Reilly, T.P., Sinclair, G.P., & McNamara, D.S. (2004). iSTART: A Web-based Reading Strategy Intervention that Improves Students' Science Comprehension. In S. D. G. Kinshuk & P. Isaías (Eds.), *nternational Conference Cognition and Exploratory Learning in Digital Age (CELDA2004)*. Lisbon, Portugal: IADIS Press.

Olston, C., Reed, B., Srivastava, U., Kumar, R., & Tomkins, A. (2008). Pig latin: a not-so-foreign language for data processing. In *ACM SIGMOD Int. Conf. on Management of Data (SIGMOD '08)* (pp. 1099–1110). Vancouver, Canada: ACM.

Oltramari, A., Gangemi, A., Guarino, N., & Masolo, C. (2002). Restructuring WordNet's Top-Level: The OntoClean approach. In *OntoLex'2 Workshop, Ontologies and Lexical Knowledge Bases (LREC 2002)* (pp. 17–26). Las Palmas, Spain.

Oprescu, B., Dascalu, M., Trausan-Matu, S., Dessus, P., & Bianco, M. (in press). Automated Assessment of Paraphrases in Pupil's Self-Explanations. *Scientific Bulletin, University Politehnica of Bucharest, Series C.*

Oprescu, B., Dascalu, M., Trausan-Matu, S., Rebedea, T., Dessus, P., & Bianco, M. (2012). Analiza automata a auto-explicatiilor. *Revista Romana de Interactiune Om-Calculator, 5*(2), 71–76.

Ortiz-Arroyo, D. (2009). Discovering Sets of Key Players in Social Networks. In A. Abraham, A. E. Hassanien & V. Snáše (Eds.), *Computational Social Network Analysis: Trends, Tools and Research Advances* (pp. 27–48). London, UK: Springer.

Paavola, S., Lipponen, L., & Hakkarainen, K. (2004). Models of innovative knowledge communities and three metaphors of learning. *Review of Educational Research, 74*(4), 557–576.

Page, E. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238–243.

Page, E. (1968). Analyzing student essays by computer. *International Review of Education, 14*(2), 210–225.

Page, L. (2001). USA Patent No. 6,285,999: USPTO.

Paris, S., & Paris, A. (2001). Classroom Applications of Research on Self-Regulated Learning. *Educational Psychologist, 36*(2), 89–101.

Pata, K., & Sarapuu, T. (2003). Meta-communicative regulation patterns of expressive modeling on whiteboard tool. In A. Rossett (Ed.), *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2003* (pp. 1126–1129). Chesapeake, VA: AACE.

Patel, A.D., & Daniele, J.R. (2003). An empirical comparison of rhythm in language and music. *Cognition, 87*(1), B35–B45.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::Similarity - Measuring the Relatedness of Concepts. In *Nineteenth National Conference on Artificial Intelligence (AAAI-04)* (pp. 1024–1025). San Jose, CA.

Petersen, S.E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language, 23*, 89–106.

Pfeffer, P. (1989). Les pharmacies des éléphants. In P. Pfeffer (Ed.), *Vie et mort d'un géant : l'éléphant d'Afrique* (pp. 135). Paris, France: Flammarion.

Pinheiro, C.A.R. (2011). *Social Network Analysis in Telecommunications*. Hoboken, NJ: John Wiley & Sons.

References

Porter, M., & Boulton, R. (2002). Snowball. Retrieved from http://snowball.tartarus.org/

Powers, D.E., Burstein, J., Chodorow, M., Fowles, M.E., & Kukich, K. (2001). Stumping e-rater®: Challenging the validity of automated essay scoring. Princeton, NJ: Educational Testing Service.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P. (2007). Support Vector Machines *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). New York, NY: Cambridge University Press.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., & Manning, C.D. (2010). A Multi-Pass Sieve for Coreference Resolution. In *Conference on Empirical Methods in Natural Language Processing (EMNLP '10)* (pp. 492–501). Cambridge, MA: ACL.

Rebedea, T. (2012). *Computer-Based Support and Feedback for Collaborative Chat Conversations and Discussion Forums.* (Doctoral dissertation), University Politehnica of Bucharest, Bucharest, Romania.

Rebedea, T., Dascalu, M., Trausan-Matu, S., Armitt, G., & Chiru, C.G. (2011). Automatic Assessment of Collaborative Chat Conversations with PolyCAFe. In C. D. Kloos, D. Gillet, R. M. Crespo García, F. Wild & M. Wolpers (Eds.), *Towards Ubiquitous Learning - 6th European Conference of Technology Enhanced Learning (EC-TEL 2011)* (pp. 299–312). Palermo, Italy: Springer.

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. In M. Wolpers, P. Kirschner, M. Scheffel, S. Lindstaedt & V. Dimitrova (Eds.), *Sustaining TEL: From Innovation to Learning and Practice - 5th European Conference on Technology Enhanced Learning (EC-TEL 2010)* (pp. 420–425). Barcelona, Spain: Springer.

Rebedea, T., Dascalu, M., Trausan-Matu, S., & Chiru, C.G. (2011). Automatic Feedback and Support for Students and Tutors Using CSCL Chat Conversations. In S. Trausan-Matu (Ed.), *First International K-Teams Workshop on Semantic and Collaborative Technologies for the Web* (pp. 20–33). Bucharest, Romania: Politehnica Press.

Rebedea, T., & Trausan-Matu, S. (2009). Computer-assisted evaluation of CSCL chat conversations. In *9th Int. Conf. on Computer supported collaborative learning (CSCL'09)* (pp. 183–185). Rhodes, Greece: ISLS.

Renaissance Learning. (2011). Matching Books to Students: How to Use Readability Formulas and Continuous Monitoring to Ensure Reading Success. Wisconsin Rapids, WI: Renaissance Learning, Inc.

Renaissance Learning. (2012a). ATOS vs. Lexile: Which Readability Formula Is Best? Wisconsin Rapids, WI: Renaissance Learning, Inc.

Renaissance Learning. (2012b). *Star Reading: Technical Manual.* Wisconsin Rapids, WI: Renaissance Learning, Inc.

Resnik, P. (1995). Using information content to evaluate semantic similarity. In *14th International Joint Conference on Artificial Intelligence* (pp. 448–453). Montreal, Canada: Morgan Kaufmann.

Rishel, T., Perkins, A.L., Yenduri, S., & Zand, F. (2006). *Augmentation of a Term/Document Matrix with Part-ofSpeech Tags to Improve Accuracy of Latent Semantic Analysis.* Paper presented at the 5th WSEAS Int. Conf. on Applied Computer Science, Hangzhou, China.

References

Rosé, C.P., Wang, Y.C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning. *International Journal of Computer Supported Collaborative Learning, 3*(3), 237–271.

Rosenblatt, F. (1957). The Perceptron-a perceiving and recognizing automaton. Buffalo, NY: Cornell Aeronautical Laboratory.

Ruwet, N. (1972). *Langage, musique, poésie*. Paris, France: Éditions du Seuil.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika, 31*, 581–603.

Sagot, B. (2008). WordNet Libre du Francais (WOLF). Paris: INRIA. Retrieved from http://alpage.inria.fr/~sagot/wolf.html

Sagot, B., & Darja, F. (2008). Building a free French wordnet from multilingual resources. In *Ontolex 2008*. Marrakech, Maroc.

Salazar Orvig, A. (1999). *Les mouvements du discours: Style, références et dialogue dans des entretiens cliniques*. Paris, France: L'Harmattan.

Sanders, T., & Noordman, L. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes, 29*(1), 37–60.

Sarmiento, J., Trausan-Matu, S., & Stahl, G. (2005). Dialogical perspectives on narratives in collaborative mathematics problem-solving. In *International Symposium on Organizational Learning and Knowledge Work Management (OL-KWM 2005)* (pp. 88–99). Bucharest, Romania.

Sawyer, R.K. (1992). Improvisational creativity: An analysis of jazz performance. *Creativity Research Journal, 5*(3), 253–263.

Sawyer, R.K. (2003). *Group Creativity: Music, Theater, Collaboration*. Mahwah, NJ and London: Lawrence Erlbaum Associates.

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.

Scardamalia, M. (2004). CSILE/Knowledge Forum *In education and Technology: An encyclopedia* (pp. 183–192). Santa Barbara: ABC-CLIO.

Schiffrin, D. (1987). *Discourse Markers*. London, UK: Cambridge University Press.

Schmidt, A.P., & Stone, T.K.M. (2013). *Detection of Topic Change in IRC Chat Logs*. Retrieved from http://www.flwyd.dhs.org/school/ircsegmentation.pdf

Schmidt, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Int. Conf. on New Methods in Language Processing* (pp. 44–49). Manchester, UK: Citeseer.

Schmidt, H. (1995). TreeTagger - a language independent part-of-speech tagger. Stuttgart, Germany: Institute for Computational Linguistics, University of Stuttgart.

Schulze, M. (2010). Measuring textual complexity in student writing. In *American Association of Applied Linguistics (AAAL 2010)*. Atlanta, GA.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

References

Sfard, A. (2000). On Reform Movement and the Limits of Mathematical Discourse. *Mathematical Thinking and Learning, 2*(3), 157–189.

Shannon, C.E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, 27*, 379–423 & 623–656.

Shannon, C.E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal, 30*, 50–64.

Sheehan, K.M., Kostin, I., & Futagi, Y. (2007). SourceFinder: A construct-driven approach for locating appropriately targeted reading comprehension source texts. In *Workshop of the International Speech Communication Association, Special Interest Group on Speech and Language Technology in Education*. Farmington, PA.

Sheehan, K.M., Kostin, I., Futagi, Y., & Flor, M. (2010). Generating automated text complexity classifications that are aligned with published text complexity standards. Princeton, NJ: Educational Testing Service.

Silber, G., & McCoy, K. (2003). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics - Summarization, 28*(4), 487–496.

Slotnick, H. (1972). Toward a theory of computer essay grading. *Journal of Educational Measurement, 9*(4), 253–263.

Snow, R., Prakash, S., Jurafsky, D., & Ng, A.Y. (2007). Learning to Merge Word Senses. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)* (pp. 1005–1014). Prague, Czech Republic: ACL.

Stahl, G. (2006a). Analyzing and designing the group cognition experience. *International Journal of Cooperative Information Systems, 15*(2), 157–178.

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

Stahl, G. (2006c). Sustaining group cognition in a math chat environment. *Research and Practice in Technology Enhanced Learning, 1*(2), 85–113.

Stahl, G. (2009a). *Studying Virtual Math Teams*. New York, NY: Springer.

Stahl, G. (2009b). The VMT vision. In G. Stahl (Ed.), *Studying Virtual Math Teams*. New York, NY: Springer.

Stenner, A.J. (1996). Measuring reading comprehension with the Lexile Framework. Durham, NC: MetaMetrics, Inc.

Stenner, A.J., Koons, H.H., & Swartz, C.W. (2009). Closing the text complexity gap: Reconceptualizing the text complexity continuum. Durham, NC: MetaMetrics, Inc.

Stent, A.J., & Allen, J.F. (2000). Annotating Argumentation Acts in Spoken Dialogue. Rochester, New York, NY: University of Rochester. Computer Science Department.

Storrer, A. (2002). Coherence in text and hypertext. *Document Design, 3*(2), 156–168.

Strijbos, J.W. (2011). Assessment of (Computer-Supported) Collaborative Learning. *IEEE Transactions on Learning Technologies, 4*(1), 59–73.

References

Suthers, D., Dwyer, N., Medina, R., & Vatrapu, R. (2007). A framework for eclectic analysis of collaborative interaction. In C. A. Chinn, G. Erkens & S. Puntambekar (Eds.), *7th Int. Conf, on Computer Supported Collaborative Learning (CSCL'07)* (pp. 695–704). New Brunswick, NJ, USA: ISLS.

Tapiero, I. (2007). *Situation models and levels of coherence*. Mahwah, NJ: Erlbaum.

Teplovs, C. (2008). The Knowledge Space Visualizer: A Tool for Visualizing Online Discourse. In *Workshop on A Common Framework for CSCL Interaction Analysis, ICLS 2008*. Utrecht, Netherland.

Thusoo, A., Sarma, J.S., Jain, N., Shao, Z., Chakka, P., Anthony, S., . . . Murthy, R. (2009). Hive: a warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment, 2*(2), 1626–1629

Toulmin, S.E. (1958). *The Uses of Arguments*. Cambridge: Cambridge University Press.

Toutanova, K., Klein, D., Manning, C.D. , & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *HLT-NAACL 2003* (pp. 252–259). Edmonton, Canada: ACL.

Toutanova, K., & Manning, C.D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*. Hong Kong: ACL.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language, 24*, 612–630.

Trausan-Matu, S. (2010a). Automatic Support for the Analysis of Online Collaborative Learning Chat Conversations. In P. M. Tsang, S. K. S. Cheung, V. S. K. Lee & R. Huang (Eds.), *3rd Int. Conf. on Hybrid Learning* (pp. 383–394). Beijing, China: Springer.

Trausan-Matu, S. (2010b). Computer Support for Creativity in Small Groups using chats. *Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information, 3*(2), 81–90.

Trausan-Matu, S. (2010c). The Polyphonic Model of Hybrid and Collaborative Learning. In F. Wang, L., J. Fong. & R. C. Kwan (Eds.), *Handbook of Research on Hybrid Learning Models: Advanced Tools, Technologies, and Applications* (pp. 466–486). Hershey, NY: Information Science Publishing.

Trausan-Matu, S. (2011). Experiencing, Conducting, Designing and Evaluating Polyphony in CSCL Chats. In H. Spada, G. Stahl, N. Miyake & N. Law (Eds.), *9th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2011)* (pp. 906–907). Hong Kong, China: ISLS.

Trausan-Matu, S. (2013). *From Two-Part Inventions for Three Voices, to Fugues and Creative Discourse Building in CSCL Chats*. Unpublished manuscript.

Trausan-Matu, S. (in press). Collaborative and Differential Utterances, Pivotal Moments, and Polyphony. In D. Suthers, K. Lund, C. P. Rosé & N. Law (Eds.), *Productive multivocality*. New York, NY: Springer.

Trausan-Matu, S., Chiru, C.G., & Bogdan, R. (2004). Identificarea actelor de vorbire in dialogurile purtate pe chat. In S. Trausan-Matu & C. Pribeanu (Eds.), *Conferinta Nationala de Interactiune Om-Calculator RoCHI 2004* (pp. 206–214). Bucharest, Romania: Printech.

Trausan-Matu, S., Dascalu, M., & Dessus, P. (2012). Textual Complexity and Discourse Structure in Computer-Supported Collaborative Learning. In S. A. Cerri, W. J. Clancey, G. Papadourakis & K. Panourgia (Eds.), *11th Int. Conf. on Intelligent Tutoring Systems (ITS 2012)* (pp. 352–357). Chania, Grece: Springer.

References

Trausan-Matu, S., Dascalu, M., & Rebedea, T. (2012). A system for the automatic analysis of Computer-Supported Collaborative Learning chats. In C. Giovannella, D. G. Sampson & I. Aedo (Eds.), *12th IEEE Int. Conf. on Advanced Learning Technologies (ICALT 2012)* (pp. 95–99). Rome, Italy: IEEE.

Trausan-Matu, S., Dessus, P., Lemaire, B., Mandin, S., Villiot-Leclercq, E., Rebedea, T., . . . Zampa, V. (2008). Deliverable D5.1 LTfLL – Support and feedback design. Heerlen: OUNL, Research report of the LTfLL Project.

Trausan-Matu, S., Dessus, P., Rebedea, T., Loiseau, M., Dascalu, M., Mihaila, D., . . . Dulceanu, A. (2011). Deliverable D5.3 LTfLL – Learning support and feedback. Heerlen: OUNL.

Trausan-Matu, S., Dessus, P., Rebedea, T., Mandin, S., Villiot-Leclercq, E., Dascalu, M., . . . Graziani, E. (2010). Deliverable D5.2 LTfLL – Learning support and feedback. Heerlen: OUNL.

Trausan-Matu, S., & Rebedea, T. (2009). Polyphonic Inter-Animation of Voices in VMT. In G. Stahl (Ed.), *Studying Virtual Math Teams* (pp. 451–473). Boston, MA: Springer.

Trausan-Matu, S., & Rebedea, T. (2010, March 21-27 2010). A Polyphonic Model and System for Inter-animation Analysis in Chat Conversations with Multiple Participants. In A. F. Gelbukh (Ed.), *11th Int. Conf. Computational Linguistics and Intelligent Text Processing (CICLing 2010)* (pp. 354–363). Iasi, Romania: Springer.

Trausan-Matu, S., Rebedea, T., & Dascalu, M. (2010). Analysis of discourse in collaborative Learning Chat Conversations with Multiple Participants. In D. Tufis & C. Forascu (Eds.), *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian* (pp. 313–330). Bucharest, Romania: Editura Academiei.

Trausan-Matu, S., Rebedea, T., Dragan, A., & Alexandru, C. (2007). Visualisation of learners' contributions in chat conversations. In J. Fong & F. L. Wang (Eds.), *Blended learning* (pp. 217–226). Singapour: Pearson/Prentice Hall.

Trausan-Matu, S., & Stahl, G. (2007). Polyphonic inter-animation of voices in chats. In *CSCL'07 Workshop on Chat Analysis in Virtual Math Teams*. New Brunwick, NJ: ISLS.

Trausan-Matu, S., Stahl, G., & Sarmiento, J. (2006). Polyphonic Support for Collaborative Learning. In Y. A. Dimitriadis, I. Zigurs & E. Gómez-Sánchez (Eds.), *Groupware: Design, Implementation, and Use, 12th International Workshop (CRIWG 2006)* (pp. 132–139). Medina del Campo, Spain: Springer.

Trausan-Matu, S., Stahl, G., & Sarmiento, J. (2007). Supporting polyphonic collaborative learning. *E-service Journal, 6*(1), 58–74.

Trausan-Matu, S., Stahl, G., & Zemel, A. (2005). Polyphonic Inter-animation in Collaborative Problem Solving Chats. Philadelphia: Drexel University.

Tutte, W.T. (2001). *Graph Theory*. Cambridge: Cambridge University Press.

Upton, G., & Cook, I. (2008). *A Dictionary of Statistics*. Oxford: Oxford University Press.

van Dijk, T.A. (1977). Coherence *Text and Context: Exploration in the Semantics and Pragmatics of Discourse* (pp. 93–129). London, UK: Longman.

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

Vidal, N. (1984). *Miguel de la faim*. Paris: Rageot.

References

Vitale, M.R., & Romance, N.R. (2007). A knowledge-based framework for unifying content-area reading comprehension and reading comprehension strategies. In D. S. McNamara (Ed.), *Reading comprehension strategies* (pp. 73–104). Mahwah: Erlbaum.

Vohs, K.D., & Baumeister, R.F. (Eds.). (2011). *Handbook of Self-Regulation: Research, theory, and applications.* New York, NY & London: The Guildford Press.

Voorhees, E.M. (1994). Query expansion using lexical-semantic relations. In W. B. Croft & C. J. v. Rijsbergen (Eds.), *17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94)* (pp. 61–69). Dublin, Ireland: Springer.

Vygotsky, L.S. (1978). *Mind in Society.* Cambridge, MA: Harvard University Press.

Wang, T., & Hirst, G. (2011). Refining the Notions of Depth and Density in WordNet-based Semantic Similarity Measures. In *Conference on Empirical Methods in Natural Language Processing (EMNLP '11)* (pp. 1003–1011). Edinburgh, Scotland, UK: ACL.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press.

Wegerif, R. (2005). A dialogical understanding of the relationship between CSCL and teaching thinking skills. In T. Koschmann, D. Suthers & T. W. Chan (Eds.), *Conf. on Computer Supported Collaborative Learning 2005: The Next 10 Years! (CSCL'05).* Taipei, Taiwan: ISLS.

Wegerif, R. (2006). A Dialogical Understanding of the Relationship between CSCL and Teaching Thinking Skills. *International Journal of Computer-Supported Collaborative Learning, 1*(1), 143–157.

Weltzer-Ward, L., Baltes, B., & Lynn, L.K. (2009). Assessing quality of critical thought in online discussion. *Campus-Wide Information Systems, 26*(3), 168–177.

Wenger, E. (1999). *Communities of practice. Learning, meaning, and identity (Learning in Doing: Social, Cognitive and Computational Perspectives).* Cambridge: Cambridge University Press.

Wertsch, J. (1998). *Mind as action.* Oxford: Oxford University Press.

White, D., & Harary, F. (2001). The Cohesiveness of Blocks in Social Networks: Node Connectivity and Conditional Density. *Sociological Methodology, 31*(1), 305–359.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing, 4*(2), 65–85.

Widdowson, H.G. (1978). *Teaching Language as Communication.* Oxford: Oxford University Press.

Wiemer-Hastings, P., & Zipitria, I. (2000). Rules for syntax, vectors for semantics. In *22nd Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Williams, M. (2002). *Wittgenstein, Mind and Meaning: Towards a Social Conception of Mind.* New York, NY: Routledge.

Winters, F.I., Greene, J.A., & Costich, C.M. (2008). Self-Regulation of Learning within Computer-based Learning Environments: A Critical Analysis. *Educational Psychology Review, 20*(4), 429–444.

Witter, D., & Berry, M.W. (1998). Downdating the latent semantic indexing model for conceptual information retrieval. *The Computer Journal, 41*, 589–601.

Wolfe, M.B.W., Magliano, J.P., & Larsen, B. (2005). Causal and semantic relatedness in discourse understanding and representation. *Discourse Processes, 39*(2-3), 165–187.

References

Wresch, W. (1993). The imminence of grading essays by computer—25 years later. *Computers and Composition, 10*(2), 45–58.

Wu, X., Kumar, V., Quinlan, J. Ross, Ghosh, J., Yang, Q., Motoda, H., . . . Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1–37.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics, ACL '94* (pp. 133–138). New Mexico, USA: ACL.

Yeh, J.-F., Wu, C.-H., & Chen, M.-J. (2008). Ontology-based speech act identification in a bilingual dialog system using partial pattern trees. *Journal of the American Society for Information Science and Technology, 59*(5), 684–694.

Zampa, V., & Dessus, P. (2012). Validating a computer-based tutor that promotes Self-Regulated Writing-to-Learn. In Y. Psaromiligkos, T. Spyridakos & S. Retalis (Eds.), *Evaluation in e-learning* (pp. 159–172). New York, NY: Nova Publishers.

Zaromb, F.M., Howard, M.W., Dolan, E.D., Sirotin, Y.B., Tully, M., Wingfield, A., & Kahana, M.J. (2006). Temporal Associations and Prior-List Intrusions in Free Recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 792–804.

Zeno, S.M., Ivens, S.H., Millard, R.T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates, Inc.

Zhang, X., Mostow, J., Duke, N., Trotochaud, C., Valeri, J., & Corbett, A.T. (2008). Mining Free-form Spoken Responses to Tutor Prompts. In R. S. J. d. Baker, T. Barnes & J. E. Beck (Eds.), *1st Int. Conf. on Educational Data Mining (Educational Data Mining 2008)* (pp. 234–241). Montreal, Canada: http://www.educationaldatamining.org/EDM2008/.

Zwaan, R.A., & Singer, M. (2003). Text comprehension. In A. C. Graesser, M. A. Gernsbacher & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 83–121). Mahwah, NJ: Erlbaum.

Appendixes – *ReaderBench* Workflows, Print-screens and Input Examples

As it can be observed from Figure 70 and Figure 71, *ReaderBench* consists of three main tabs, each focused on specific functionalities: 1/ *document* assessment, both for general texts and conversations, 2/ *verbalization* assessment and 3/ *textual complexity* corpus training and evaluation. The conventions of all the workflows are the following: *arcs* denote possible transition, whereas *rectangles* signify actions (flat bottom rectangle) or other interfaces depicted as figures within the thesis (wavy bottom rectangle).

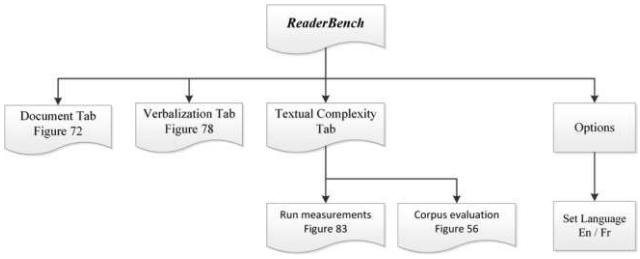


Figure 70. *ReaderBench* General workflow.

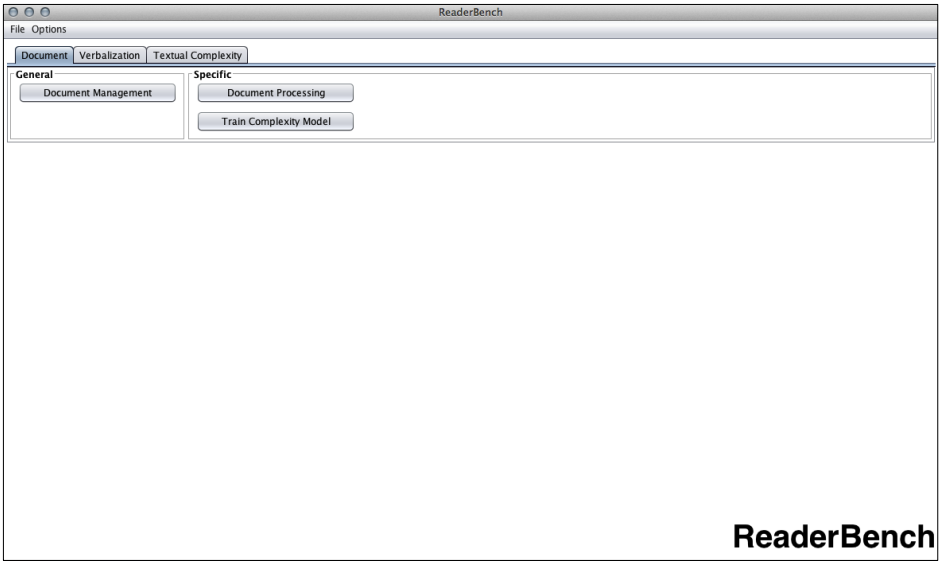


Figure 71. *ReaderBench* Main user interface.

Appendix A – Document Workflow and Additional Print-screens

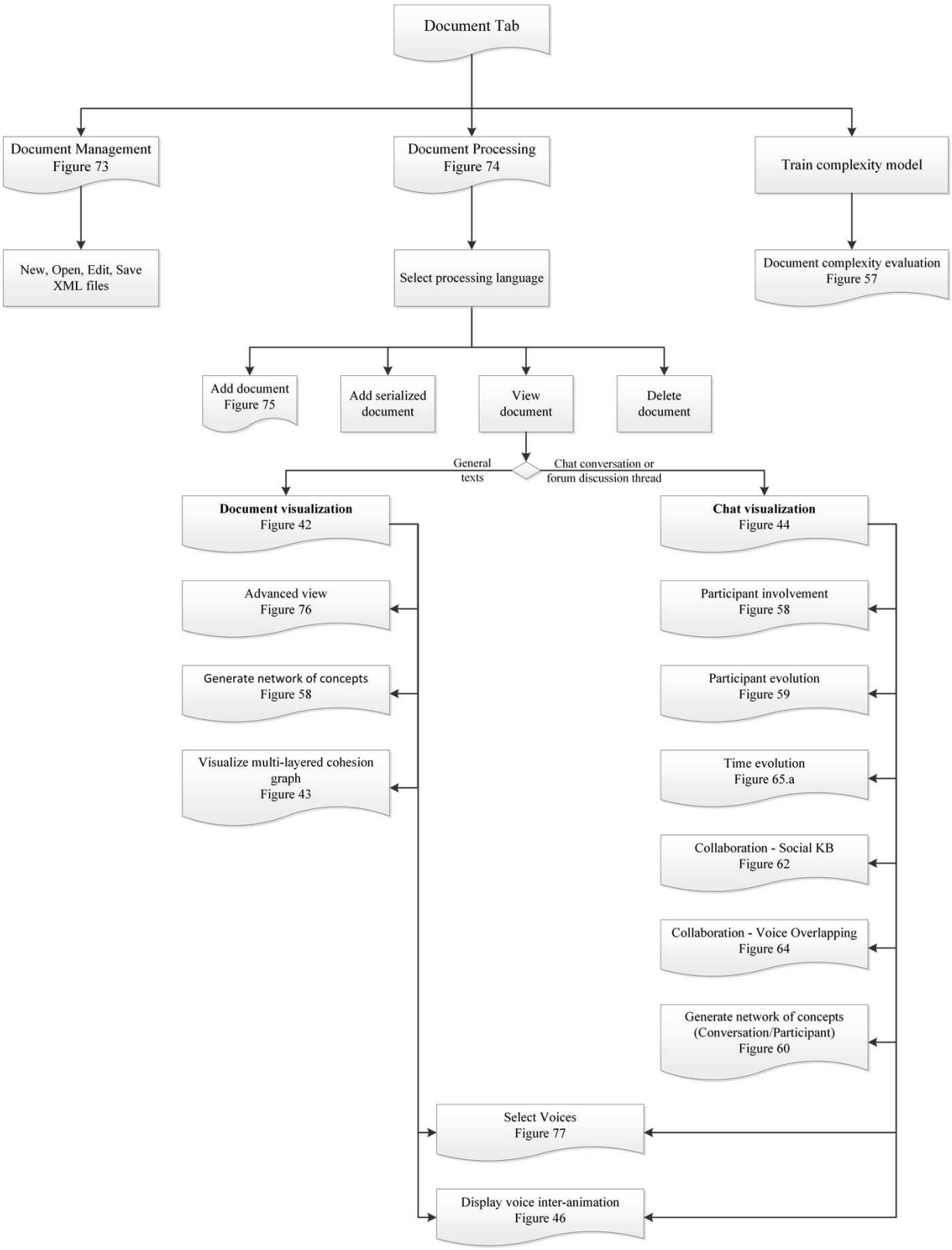


Figure 72. *ReaderBench* Document workflow.

Figure 73. *ReaderBench* Document management interface.

Enables the possibilities for the tutor to create new, load, edit and save texts (reading materials) in corresponding XML format, with all recommended fields, including the possibility to define verbalization breakpoints that are automatically considered when generating new learner self-explanations (see Figure 79)

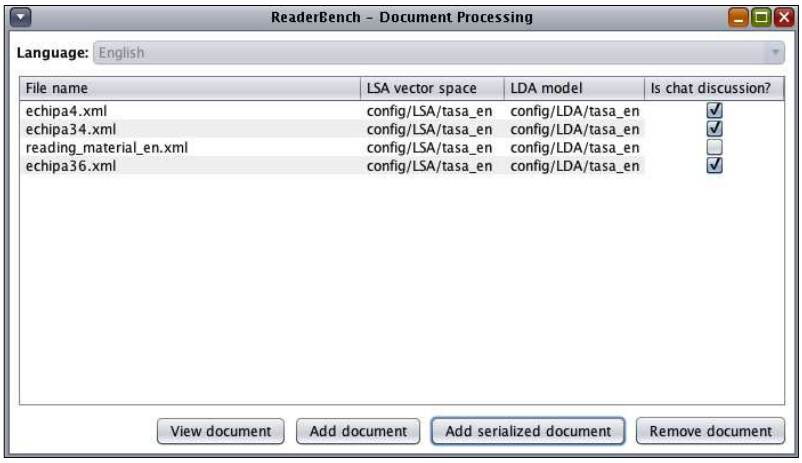


Figure 74. *ReaderBench* Document processing interface.

Allows users to add, remove or visualize a loaded document (conversation or general text). Serialized documents are pre-computed documents that are saved as serialized Java objects and can be easily recovered in order to eliminate the processing time required for a new document

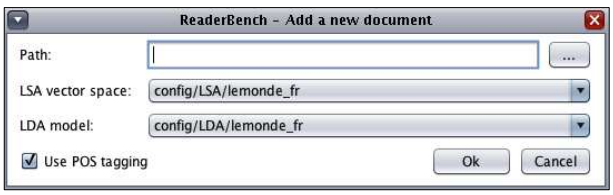
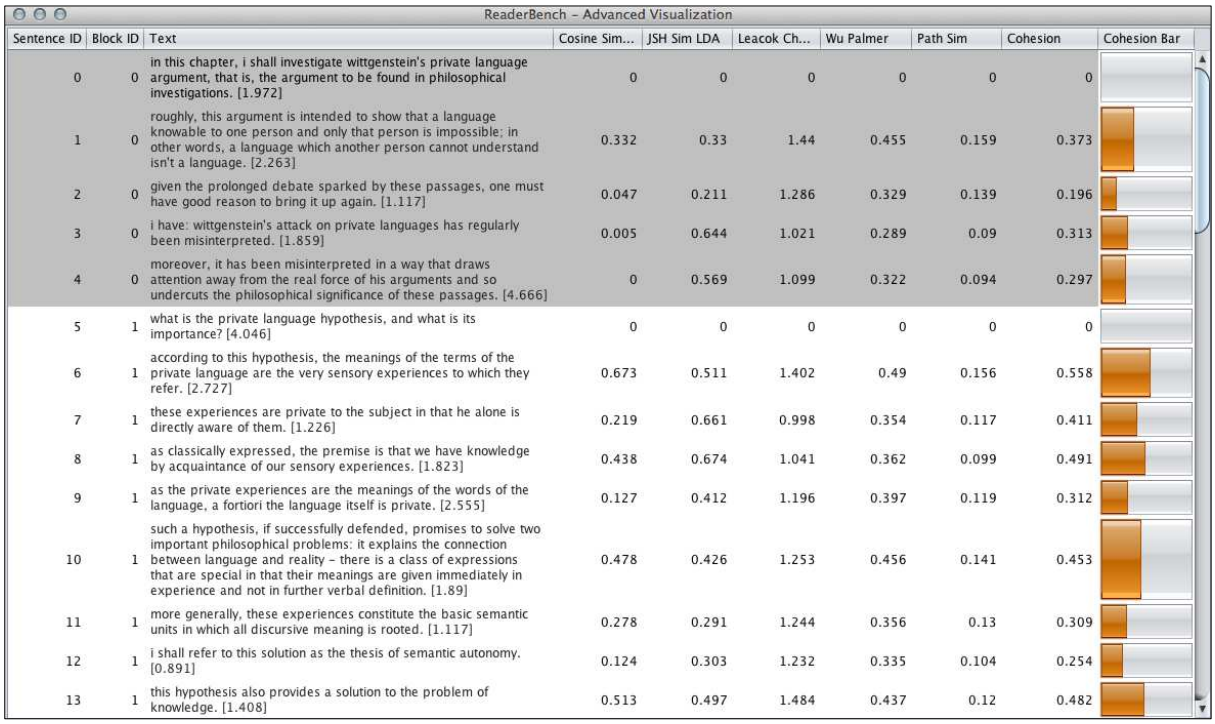


Figure 75. *ReaderBench* Interface for adding a new document for processing.



Sentence ID	Block ID	Text	Cosine Sim...	JSH Sim LDA	Leacok Ch...	Wu Palmer	Path Sim	Cohesion	Cohesion Bar
0	0	in this chapter, i shall investigate wittgenstein's private language argument, that is, the argument to be found in philosophical investigations. [1.972]	0	0	0	0	0	0	
1	0	roughly, this argument is intended to show that a language knowable to one person and only that person is impossible; in other words, a language which another person cannot understand isn't a language. [2.263]	0.332	0.33	1.44	0.455	0.159	0.373	
2	0	given the prolonged debate sparked by these passages, one must have good reason to bring it up again. [1.117]	0.047	0.211	1.286	0.329	0.139	0.196	
3	0	i have: wittgenstein's attack on private languages has regularly been misinterpreted. [1.859]	0.005	0.644	1.021	0.289	0.09	0.313	
4	0	moreover, it has been misinterpreted in a way that draws attention away from the real force of his arguments and so undercuts the philosophical significance of these passages. [4.666]	0	0.569	1.099	0.322	0.094	0.297	
5	1	what is the private language hypothesis, and what is its importance? [4.046]	0	0	0	0	0	0	
6	1	according to this hypothesis, the meanings of the terms of the private language are the very sensory experiences to which they refer. [2.727]	0.673	0.511	1.402	0.49	0.156	0.558	
7	1	these experiences are private to the subject in that he alone is directly aware of them. [1.226]	0.219	0.661	0.998	0.354	0.117	0.411	
8	1	as classically expressed, the premise is that we have knowledge by acquaintance of our sensory experiences. [1.823]	0.438	0.674	1.041	0.362	0.099	0.491	
9	1	as the private experiences are the meanings of the words of the language, a fortiori the language itself is private. [2.555]	0.127	0.412	1.196	0.397	0.119	0.312	
10	1	such a hypothesis, if successfully defended, promises to solve two important philosophical problems: it explains the connection between language and reality – there is a class of expressions that are special in that their meanings are given immediately in experience and not in further verbal definition. [1.89]	0.478	0.426	1.253	0.456	0.141	0.453	
11	1	more generally, these experiences constitute the basic semantic units in which all discursive meaning is rooted. [1.117]	0.278	0.291	1.244	0.356	0.13	0.309	
12	1	i shall refer to this solution as the thesis of semantic autonomy. [0.891]	0.124	0.303	1.232	0.335	0.104	0.254	
13	1	this hypothesis also provides a solution to the problem of knowledge. [1.408]	0.513	0.497	1.484	0.437	0.12	0.482	

Figure 76. *ReaderBench* Document advanced visualization.
An advanced view available only for documents that presents cohesion-based scores using different semantic measures for adjacent sentences of the same paragraph

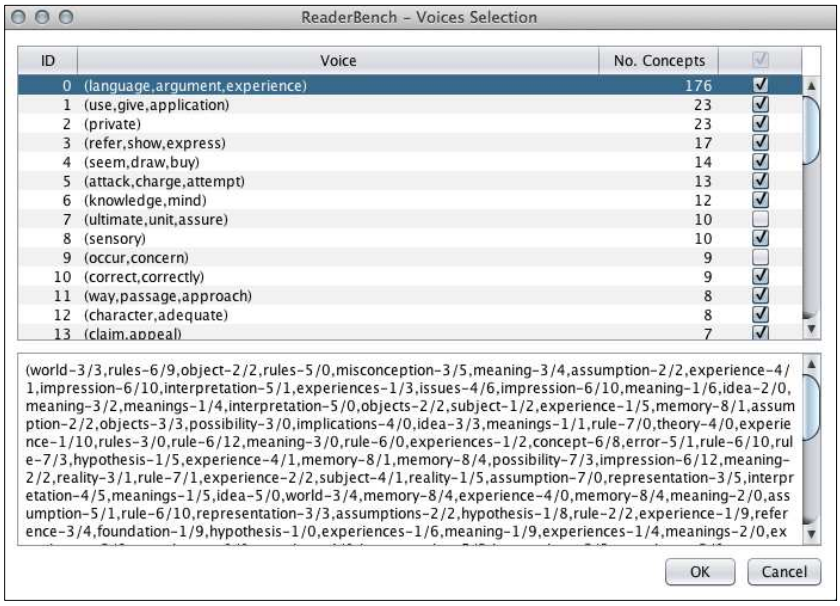


Figure 77. *ReaderBench* Voice selection interface.
This interface enables the user to manually select the voices (semantic word chains) of interest, ordered in descending order of the number of comprised concepts, later to be used for displaying voice inter-animation (at least one voice must be selected) (see Figure 46). By default, as the number of overall voices can become cumbersome, all voices are deselected. A voice is displayed as a tuple of the 3 most frequently occurring word lemmas within the semantic chain. The lower part of the interface displays the entire selected semantic chain, consisting of flectional word forms followed by paragraph ID/sentence ID (general texts) or utterance ID/sentence ID (conversations or forum discussion threads)

Appendix B – Verbalization Workflow and Additional Print-screens

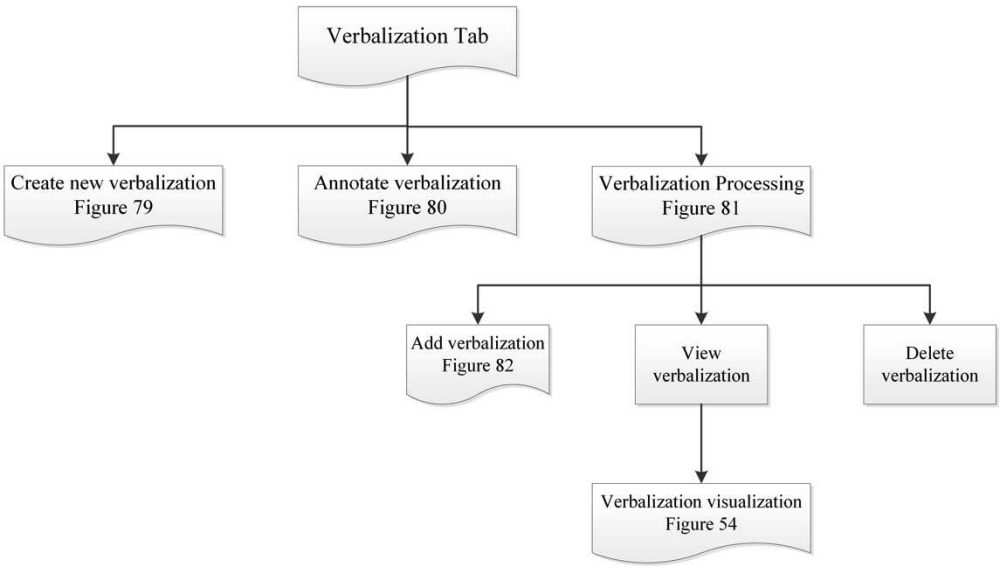


Figure 78. *ReaderBench* Verbalization workflow.

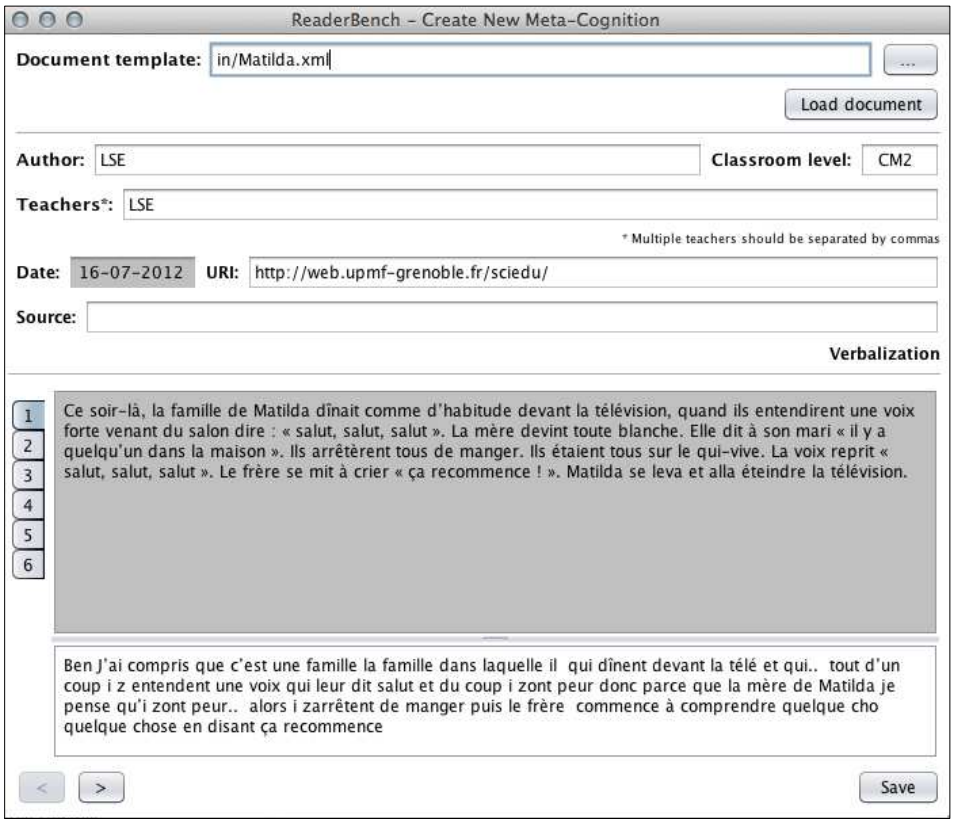


Figure 79. *ReaderBench* Interface for creating new self-explanations.
An interface designed for learners, enabling them to add their self-explanations in an intuitive manner, after each corresponding chunk of text from the original reading material

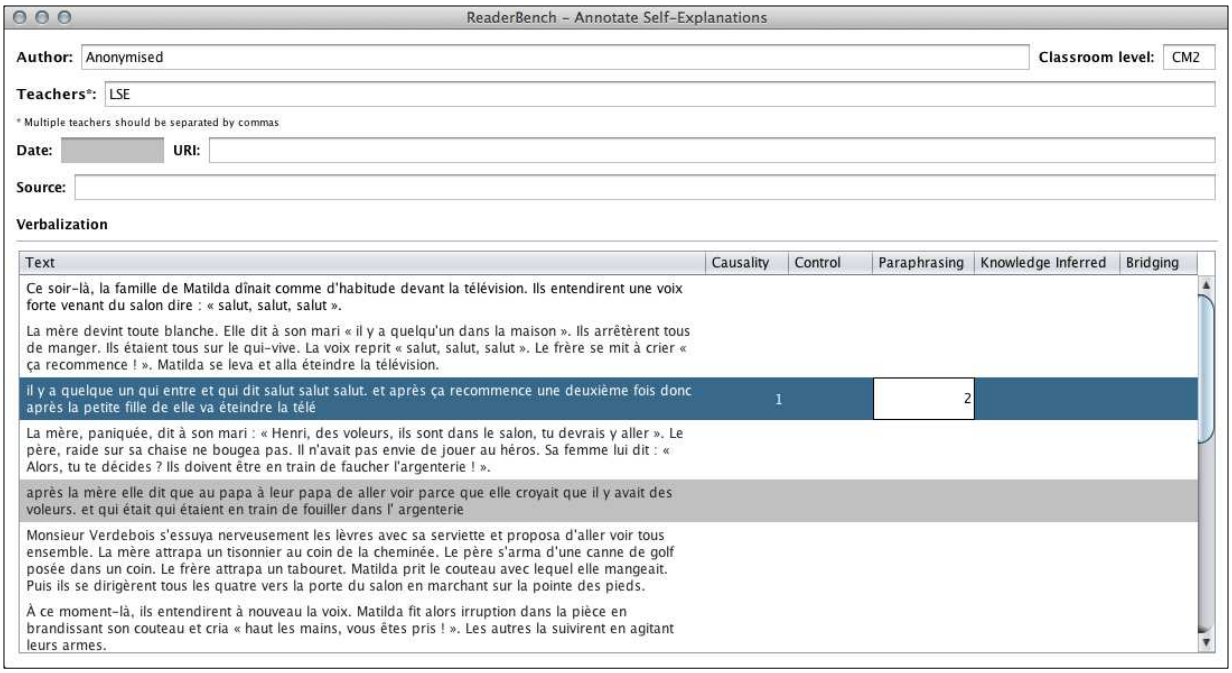


Figure 80. *ReaderBench* Interface for manually annotating self-explanations.
Tutors are granted the possibility to load learner self-explanations and annotate them with the correspondingly identified reading strategies

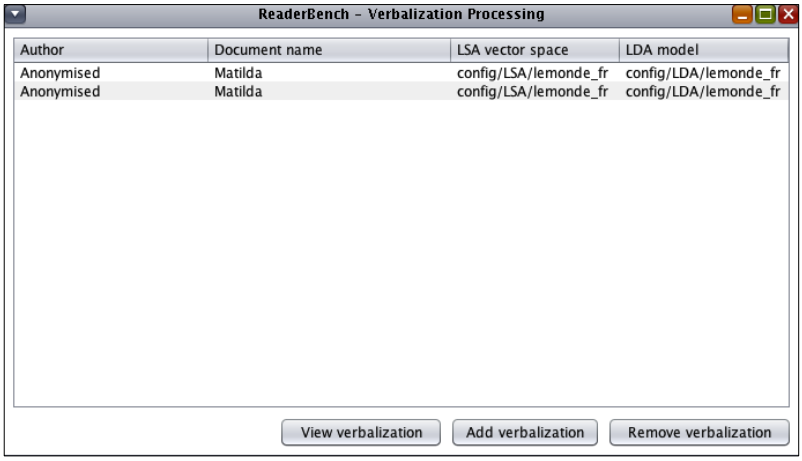


Figure 81. *ReaderBench* Verbalization processing interface.

Add, view and deleted operations on learner verbalizations. At least one document must be loaded

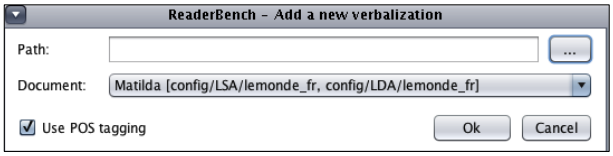


Figure 82. *ReaderBench* Interface for adding a new verbalization for processing.

Appendix C – Textual Complexity Additional Print-screen and EA/AA Scores

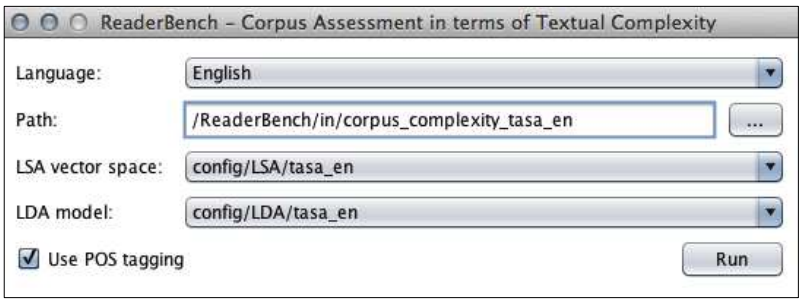


Figure 83. *ReaderBench* Corpus textual complexity assessment interface.
Determines all textual complexity factors for all the documents within the training corpora. All the documents for a specific complexity class are included as separate XML files within a folder named "class<ID>". In addition, the user can select the LSA and LDA semantic models he/she wants to use in the assessment process and whether part-of-speech tagging should be applied on the corpus or not

The detailed evaluation results in terms of all complexity dimensions and all withheld textual complexity factors are presented in Table 40. The factors with a high agreement (EA ≥ 0.5 and AA ≥ .85) are marked with *italics* and "*".

Table 40. Exact Agreement (EA) and Adjacent Agreement (AA) for all evaluation factors.

Factor	*C1(%)* EA/AA	*C2(%)* EA/AA	*C3(%)* EA/AA	*C4(%)* EA/AA	*C5(%)* EA/AA	*C6(%)* EA/AA	*Avg.* EA/AA
*All Factors Combined**	.9/1	.8/1	.76/1	.7/.99	.7/1	.82/.99	.78/1
Textual Complexity Dimensions							
*Readability Factors**	.82/1	.72/1	.68/1	.57/.99	.68/.99	.78/.98	.71/.99
Fluency Factors	.74/.76	.03/.57	.18/.28	.14/.37	.19/.68	.64/.76	.32/.57
*Structure Complexity Factors**	.91/1	.76/.99	.61/1	.62/.97	.66/.99	.73/.99	.72/.99
*Diction Factors**	.85/1	.61/.97	.43/.9	.41/.86	.26/.86	.71/.86	.55/.91
Entropy Factors	.58/.68	.23/.62	.07/.3	.12/.35	.2/.72	.56/.72	.3/.56
*Balanced CAF Factors**	.87/1	.81/1	.72/1	.68/.99	.66/1	.77/.99	.75/1
*Part of Speech Complexity Factors**	.87/.99	.58/.98	.5/.95	.35/.91	.41/.89	.67/.87	.56/.93
Parsing Tree Complexity Factors	.73/.96	.45/.88	.18/.74	.35/.62	.13/.8	.65/.76	.42/.79
*Named Entity Complexity Factors**	.87/1	.59/.96	.49/.92	.44/.88	.37/.9	.69/.87	.58/.92
Co-reference Complexity Factors	.68/.9	.34/.78	.26/.64	.29/.62	.17/.79	.45/.71	.37/.74
*Word Complexity Factors**	.7/.98	.52/.98	.47/.87	.34/.88	.52/.9	.73/.95	.55/.93
Lexical Chains Factors	.62/.85	.41/.8	.21/.62	.23/.54	.13/.75	.58/.72	.36/.71

Factor	*C1(%)* EA/AA	*C2(%)* EA/AA	*C3(%)* EA/AA	*C4(%)* EA/AA	*C5(%)* EA/AA	*C6(%)* EA/AA	*Avg.* EA/AA
*Discourse Factors**	.78/.98	.49/.94	.42/.86	.38/.82	.39/.91	.69/.84	.53/.89
Readability Factors							
*Readability Flesh**	.74/.99	.49/.99	.52/.88	.29/.88	.51/.92	.78/.98	.55/.94
*Readability FOG**	.72/.95	.52/.94	.28/.87	.34/.84	.49/.94	.74/.95	.51/.91
*Readability Kincaid**	.72/1	.57/.99	.53/.95	.39/.95	.54/.97	.79/.98	.59/.97
*Number of words per sentence**	.76/.98	.62/.95	.4/.93	.38/.88	.38/.91	.68/.86	.54/.92
*Average number of syllables per word**	.73/.97	.46/.98	.52/.83	.27/.83	.47/.87	.72/.98	.53/.91
Percentage of complex words (>2 syllables)	.71/.92	.32/.91	.28/.69	.17/.71	.35/.8	.72/.95	.42/.83
Fluency Factors							
Normalized number of commas	.16/.28	.12/.39	.23/.52	.19/.53	.22/.61	.39/.57	.22/.48
Normalized number of words	.72/.75	.06/.58	.27/.31	.03/.23	.01/.54	.67/.73	.29/.52
Structure Complexity Factors							
Normalized number of blocks	.79/.95	.3/.83	.05/.23	.07/.19	.12/.74	.78/.87	.35/.63
Average block size	.76/.97	.48/.9	.3/.68	.18/.45	.26/.84	.68/.77	.44/.77
Normalized number of sentences	.7/.97	.51/.85	.15/.77	.41/.57	.08/.86	.78/.81	.44/.8
*Average sentence length**	.82/1	.6/.97	.41/.86	.4/.84	.29/.86	.67/.86	.53/.9
Diction Factors							
*Average word length**	.84/1	.63/.98	.44/.91	.41/.89	.26/.88	.66/.86	.54/.92
Entropy Factors							
Word entropy	.68/.88	.44/.91	.02/.38	.02/.08	.05/.18	.11/.16	.22/.43
Character entropy	.56/.56	.04/.53	.1/.28	.24/.32	0/.78	.71/.71	.27/.53
Balanced CAF Factors							
Lexical Diversity	.77/.88	.22/.87	.33/.46	.09/.37	.05/.38	.26/.35	.29/.55
*Lexical Sophistication**	.84/1	.78/1	.69/1	.59/.99	.58/1	.81/.98	.71/1
Syntactic Diversity	.7/.74	.11/.77	.22/.57	.25/.62	.14/.38	0/.11	.24/.53
*Syntactic Sophistication**	.85/.99	.57/.98	.49/.95	.34/.87	.36/.88	.7/.85	.55/.92
Balanced CAF	.83/.99	.61/.99	.54/.84	.16/.71	.24/.62	.5/.68	.48/.8

Factor	*C1(%)* EA/AA	*C2(%)* EA/AA	*C3(%)* EA/AA	*C4(%)* EA/AA	*C5(%)* EA/AA	*C6(%)* EA/AA	*Avg.* EA/AA
Part of Speech Complexity Factors							
*Average number of nouns**	.86/1	.57/.96	.5/.9	.35/.81	.23/.86	.66/.82	.53/.89
Average number of pronouns	.77/.89	.17/.77	.08/.29	.08/.19	.05/.53	.45/.51	.27/.53
Average number of verbs	.75/.89	.21/.71	.09/.4	.13/.31	.15/.75	.67/.76	.33/.64
Average number of adverbs	.63/.89	.3/.88	.16/.36	0/.21	.16/.29	.25/.34	.25/.5
Average number of adjectives	.77/.95	.44/.95	.38/.8	.31/.75	.22/.72	.43/.65	.42/.8
Average number of prepositions	.76/.99	.59/.93	.31/.85	.4/.71	.07/.82	.65/.7	.46/.83
Parsing Tree Complexity Factors							
Average tree depth	.65/.93	.39/.81	.2/.73	.36/.6	.1/.79	.66/.71	.39/.76
Average tree size	.75/.97	.45/.86	.18/.74	.39/.57	.08/.83	.74/.8	.43/.79
Named Entity Complexity Factors							
Total number of named entities	.63/.76	.17/.73	.12/.26	.01/.07	.03/.5	.69/.71	.28/.51
Total number of entities per document	.67/.88	.33/.68	0/.38	.26/.37	.24/.69	.37/.53	.31/.59
Total number of unique entities per document	.61/.88	.27/.63	.03/.44	.29/.34	.06/.65	.49/.52	.29/.58
Percentage of entities per document	.73/.76	.05/.61	.09/.29	.2/.44	.18/.7	.44/.56	.28/.56
Percentage of unique entities per document	.48/.78	.37/.61	.11/.51	.24/.41	.1/.65	.34/.42	.27/.56
*Average number of entities per sentence**	.85/1	.54/.98	.44/.91	.34/.79	.22/.83	.67/.82	.51/.89
Average number of unique entities per sentences	.87/1	.48/.97	.35/.77	.2/.6	.19/.64	.45/.55	.42/.75
Percentage of named entities per document	.63/.73	.11/.7	.08/.16	.05/.07	.04/.56	.72/.74	.27/.49
Average number of named entities per sentence	.74/.87	.24/.79	.19/.46	.06/.29	.13/.41	.34/.36	.28/.53
Percentage of named entities in total entities	.67/.75	.11/.72	.06/.2	.05/.16	.08/.62	.76/.81	.29/.54
Percentage of nouns in total entities	.73/.84	.09/.78	.05/.22	.04/.15	.12/.59	.68/.74	.28/.55
Percentage of nouns per document	.48/.73	.37/.67	.08/.54	.21/.44	.16/.79	.69/.83	.33/.67

Factor	*C1(%)* EA/AA	*C2(%)* EA/AA	*C3(%)* EA/AA	*C4(%)* EA/AA	*C5(%)* EA/AA	*C6(%)* EA/AA	*Avg.* EA/AA
*Average number of nouns per sentence**	.83/.98	.59/.94	.53/.95	.44/.89	.35/.91	.65/.89	.57/.93
Percentage remaining nouns per document	.55/.8	.3/.77	.21/.57	.21/.34	0/.66	.63/.63	.32/.63
Average number of remaining nouns per sentence	.85/1	.54/.97	.48/.84	.23/.68	.29/.77	.58/.8	.5/.84
Percentage of overlapping nouns per document	.48/.76	.22/.67	.03/.24	.01/.02	0/.6	.78/.8	.25/.52
Average number of overlapping nouns per sentence	.82/.85	.12/.79	.17/.33	.06/.23	.1/.55	.44/.5	.28/.54
Co-reference Complexity Factors							
Total number of co-reference chains per document	.14/.41	.26/.55	.13/.47	.19/.44	.21/.72	.38/.59	.22/.53
Average number of co-references per chain	.42/.47	.11/.43	.16/.32	.1/.31	.08/.8	.76/.82	.27/.52
Average co-reference chain span	.65/.82	.27/.65	.01/.41	.33/.36	.04/.64	.38/.41	.28/.55
Number of co-reference chains with a big span	.15/.47	.33/.58	.16/.4	.06/.38	.21/.68	.53/.72	.24/.54
Average inference distance per co-reference chain	.63/.79	.31/.84	.1/.34	0/.2	.24/.41	.16/.36	.24/.49
Number of active co-reference chains per word	.29/.53	.33/.64	.19/.57	.23/.5	.11/.62	.4/.5	.26/.56
Number of active co-reference chains per entity	.4/.68	.29/.82	.22/.54	.05/.27	.02/.36	.41/.45	.23/.52
Word Complexity Factors							
Mean distance between lemma and word stems	.6/.9	.38/.93	.3/.76	.32/.74	.43/.83	.62/.88	.44/.84
Mean distance between words and corresponding stems	.61/.85	.25/.91	.32/.72	.27/.79	.48/.82	.62/.9	.42/.83
Mean word distance in hypernym tree	.09/.31	.2/.67	.52/.73	.08/.7	.15/.35	.1/.25	.19/.5
*Mean word polysemy count**	.73/.97	.5/.98	.54/.89	.43/.86	.45/.9	.75/.97	.57/.93
*Mean word syllable count**	.76/.98	.51/.98	.56/.89	.35/.84	.47/.88	.72/.96	.56/.92

Factor	*C1(%)* EA/AA	*C2(%)* EA/AA	*C3(%)* EA/AA	*C4(%)* EA/AA	*C5(%)* EA/AA	*C6(%)* EA/AA	*Avg.* EA/AA
Lexical Chains Factors							
Average span of lexical chains	.56/.77	.29/.78	.2/.45	.05/.37	.09/.27	.1/.18	.22/.47
Maximum span of lexical chains	.77/.86	.15/.72	.12/.28	.09/.18	.04/.73	.76/.8	.32/.59
Number of lexical chains with more than 5 concepts	.66/.89	.22/.69	.05/.25	.06/.11	.01/.69	.81/.81	.3/.57
Percentage of words that are included in lexical chains with more than 5 concepts	.59/.68	.17/.7	.25/.36	.07/.35	.05/.36	.25/.34	.23/.47
Discourse Factors							
Average block score	.66/.86	.26/.66	.02/.29	.17/.25	.12/.78	.79/.84	.34/.61
Overall document score	.39/.66	.4/.73	.24/.55	.06/.47	.22/.33	.08/.25	.23/.5
Average block-document cohesion	.68/.93	.29/.72	.02/.36	.22/.25	.03/.75	.81/.83	.34/.64
Average sentence-block cohesion	.75/.84	.18/.58	.03/.35	.32/.38	.04/.79	.76/.8	.34/.62
Average inter-block cohesion	.51/.75	.22/.58	.01/.26	.1/.15	.05/.71	.81/.83	.28/.55
Average intra-block cohesion	.56/.75	.24/.62	.11/.39	.16/.38	.11/.74	.72/.79	.32/.61

Appendix D – Input Examples

Sample Document – Matilda by Dahl (2007)

Ce soir-là, la famille de Matilda dînait comme d'habitude devant la télévision, quand ils entendirent une voix forte venant du salon dire : « salut, salut, salut ». La mère devint toute blanche. Elle dit à son mari « il y a quelqu'un dans la maison ». Ils arrêtèrent tous de manger. Ils étaient tous sur le qui-vive. La voix reprit « salut, salut, salut ». Le frère se mit à crier « ça recommence ! ». Matilda se leva et alla éteindre la télévision.

<< *Verbalization breakpoint 1* >>

La mère, paniquée, dit à son mari : « Henri, des voleurs, ils sont dans le salon, tu devrais y aller ». Le père, raide sur sa chaise ne bougea pas. Il n'avait pas envie de jouer au héros.

Sa femme lui dit : « Alors, tu te décides ? Ils doivent être en train de faucher l'argenterie ! ».

<< *Verbalization breakpoint 2* >>

Monsieur Verdebois s'essuya nerveusement les lèvres avec sa serviette et proposa d'aller voir tous ensemble. La mère attrapa un tisonnier au coin de la cheminée. Le père s'arma d'une canne de golf posée dans un coin. Le frère attrapa un tabouret. Matilda prit le couteau avec lequel elle mangeait. Puis ils se dirigèrent tous les quatre vers la porte du salon en marchant sur la pointe des pieds.

À ce moment-là, ils entendirent à nouveau la voix. Matilda fit alors irruption dans la pièce en brandissant son couteau et cria « haut les mains, vous êtes pris ! ». Les autres la suivirent en agitant leurs armes.

<< *Verbalization breakpoint 3* >>

Puis, ils s'arrêtèrent pour regarder autour d'eux. Ils ne virent personne. Le père fut soulagé et dit « il n'y a pas de voleur ici ». Sa femme lui répondit d'une voix tremblante « mais Henri, je l'ai entendu, et toi aussi ». Matilda appuya la réponse de sa mère en ajoutant « je suis sûre de l'avoir entendu, il est ici quelque part ».

C'est alors que la voix s'éleva à nouveau. Ils sursautèrent tous, y compris Matilda qui jouait très bien la comédie. Ils inspectèrent la grande pièce. Ils ne trouvèrent toujours personne.

<< *Verbalization breakpoint 4* >>

Matilda dit alors que c'était un fantôme : « Le salon est hanté, je croyais que vous le saviez. Je sais que c'est le fantôme, je l'ai déjà entendu ici ». Les parents, très pâles, sortirent du salon suivis par les enfants.

<< *Verbalization breakpoint 5* >>

Plus tard, suivie de son frère, Matilda retourna dans la pièce. C'est alors qu'elle sortit du manteau de la cheminée le perroquet de leur copain Arthur. Ils éclatèrent alors de rire. Ils passèrent par la porte de derrière en emmenant l'animal avec eux. Matilda rendit son perroquet à Arthur et lui raconta la soirée. Il n'y eut plus jamais de fantôme chez les Verdebois.

<< *Verbalization breakpoint 6* >>

Sample Chat – Log of Team 4 Chat Conversation

```
<Dialog team="4" file="Team4.xml">
    <!-- predefined topics by the tutor -->
    <Topics>
        <Topic>Blog</Topic>
        <Topic>Chat</Topic>
        <Topic>Forum</Topic>
        <Topic>Wiki</Topic>
    </Topics>
    <Body>
        <Turn nickname="Participant 1">
            <Utterance genid="1" time="03.05.23" ref="0">joins the room</Utterance>
        </Turn>
        …
        <!-- students select a technology out of the 4 initially suggested -->
        <Turn nickname="Participant 1">
            <Utterance genid="18" time="03.24.37" ref="0">I will tell you why my company
            loves blogs - in fact we have a product called Blog2007</Utterance>
        </Turn>
        …
        <!-- students present arguments and debate the pros and cons of each technology -->
        <Turn nickname="Participant 4">
            <Utterance genid="27" time="03.26.02" ref="0">I think that a chat system for a
            company is much more suitable</Utterance>
        </Turn>
        …
        <Turn nickname="Participant 1">
            <Utterance genid="41" time="03.28.49" ref="0">the major problem of wiki is that
            too many people can change the content</Utterance>
        </Turn>
        <Turn nickname="Participant 1">
            <Utterance genid="42" time="03.29.00" ref="41">and so, it can be confusing
            </Utterance>
        </Turn>
        <Turn nickname="Participant 2">
            <Utterance genid="43" time="03.29.22" ref="0">well this can be observed by
            somebody and not all changes are permanent</Utterance>
        </Turn>
        …
```

```
<Turn nickname="Participant 1">
    <Utterance genid="144" time="03.51.31" ref="140">I've seen people telling how they solved catchy problems on blogs, talking techniques of programming and so on</Utterance>
</Turn>
<Turn nickname="Participant 3">
    <Utterance genid="145" time="03.51.33" ref="0">the wiki is very professional
    </Utterance>
</Turn>
<Turn nickname="Participant 1">
    <Utterance genid="146" time="03.51.48" ref="144">they were providing solutions, not talking about their personal life</Utterance>
</Turn>
…
<!--students sum up the benefits of each technology they supported throughout the conversation-->
<Turn nickname="Participant 4">
    <Utterance genid="256" time="04.21.42" ref="0">4. The spread of chat is huge because of it's low bandwidth requirement</Utterance>
</Turn>
<Turn nickname="Participant 4">
    <Utterance genid="257" time="04.22.45" ref="0">5 You can debate and get good answers, whitch can be explained right away if misunderstood</Utterance>
</Turn>
…
<!-- students combine the technologies in order to integrated environment -->
<Turn nickname="Participant 4">
    <Utterance genid="352" time="04.50.39" ref="0">using the wiki will furthermore take away from the work of the support team</Utterance>
</Turn>
<Turn nickname="Participant 3">
    <Utterance genid="353" time="04.50.42" ref="0">the forum is where this knowledge is discussed</Utterance>
</Turn>
…
<!-- end of conversation -->
<Turn nickname="Participant 1">
    <Utterance genid="395" time="04.57.43" ref="394">bye</Utterance>
</Turn>
<Turn nickname="Participant 1">
    <Utterance genid="396" time="04.57.54" ref="0">leaves the room</Utterance>
</Turn>
</Body>
</Dialog>
```

Sample Verbalization

The main reading strategies that were manually coded according to annotation methodology (see *8.1.3* Reading Strategies Identification Heuristics) were: Paraphrasing, Control, Bridging, Causality and Knowledge Inference. In addition, two other strategies were initially coded, but later on disregarded as they were insignificant as occurrences in terms of the overall self-explanations corpus: Generalization and Prediction (see Table 41).

Table 41. Self-explanations example, manually coded in correspondence with the annotation methodology used by Nardy et al. (in press).

No.	Transcript
V1	Ben J'ai compris que c'est une famille la famille dans laquelle il /???suis/ qui dînent devant la télé et qui.. tout d'un coup i z entendent une voix qui leur dit salut et du coup i zont peur donc parce que la mère de Matilda /???donc c'est qué/ je pense qu'i zont peur.. alors i zarrêtent de manger puis le frère commence à comprendre quelque cho quelque chose en disant ça recommence
V2	alors je pense je pense que c'est une famille peut-être assez riche parce qu'il y a de l'argenterie et qui pensent que ceux qui doit être riche ou qu'y a beaucoup de voleurs dans notre dans leur maison donc...
V3	donc la c'est... on sait déjà comment s'appelle la famille et puis ils racontent que là vu que le père veut pas y aller tout seul il est accompagné de toute sa famille pour... aller... voir s'y a un voleur et y a la le la parole /ça le bruit aussi ???/qui recommence *et du coup elle, la petite fille qui s'appelle Matilda commence à avoir peur* (1)*donc* elle(1) *lui* (2)dit haut les mains vous êtes pris(2)
V4	*Donc là...eee i i disent* qu'y a pas de voleur et que par contre Matilda et sa maman sont sûres de l'avoir entendu et elle dit que y commencent à sursauter et que Matilda jouait très bien la comédie donc soit on peut penser que elle fait ça pour être d'accord avec ses parents ou soit que c'est elle qui a fait une blague ou soit que c'est peut-être une pièce .. de théâtre ou quelque chose comme ça mais je pense surtout que c'est elle qui leur a fait une blague parce que... parce que... comme
V5	Donc là je confirme que c'est une blague par ce que*(rire)*...elle c'est elle qui éveille les soupçons donc... voilà
V6	Donc là on sait que c'est une blague donc elle avait pris le perroquet de son copain pour jouer un tour à ses parents (E : tu vois autre chose à ajouter ?) Ben que... Par rapport à...

Detailed Table of Contents

Author Index