ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies

Mihai Dascalu, Philippe Dessus, Stefan Trausan-Matu, Maryse Bianco, Aurélie Nardy, Mihai Dascălu, Ștefan Trăușan-Matu

ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies

Mihai Dascălu[1,2], Philippe Dessus[2,3], Ștefan Trăușan-Matu[1], Maryse Bianco[2], and Aurélie Nardy[2]

[1] Politehnica University of Bucharest, Computer Science Department, Romania
{mihai.dascalu, stefan.trausan}@cs.pub.ro
[2] LSE, Univ. Grenoble Alpes, France
[3] LIG-MeTAH, Univ. Grenoble Alpes, France
{philippe.dessus, maryse.bianco, aurelie.nardy}@upmf-grenoble.fr

Abstract. *ReaderBench* is a multi-purpose, multi-lingual and flexible environment that enables the assessment of a wide range of learners' productions and their manipulation by the teacher. *ReaderBench* allows the assessment of three main textual features: cohesion-based assessment, reading strategies identification and textual complexity evaluation, which have been object to empirical validations. *ReaderBench* covers a complete cycle, from the initial complexity assessment of reading materials, the assignment of texts to learners, the capture of metacognitions reflected in one's textual verbalizations and comprehension evaluation, therefore fostering learner's self-regulation process.

Keywords: Text Cohesion, Reading Strategies, Textual Complexity, Latent Semantic Analysis, Latent Dirichlet Allocation, Support Vector Machines

1 Introduction

In every instructional situation, reading textual materials and writing down thoughts are the core activities that represent both *causes* of learning (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.

A teacher should take care of a small number of students for better supporting learners' reading and writing, which is difficult to be carried out on a larger scale. However, assessing textual materials and verbalizations is a cognitively demanding and subjectivity-laden activity. We thus designed and implemented *ReaderBench*, a flexible computer-based environment that supports reading and writing activities of learners and of teachers in multiple educational scenarios.

The following section details some of the main predictors of reading comprehension, leading to the introduction of *ReaderBench*. The third section is centered on the analysis of textual cohesion, considered central within discourse analysis. Then we

shift the point of interest towards reading strategies and assessing textual complexity. Each of the three latter sections is accompanied by a validation with *ReaderBench*.

2 Core Predictors of Reading Comprehension

Expert readers are strategic readers. They monitor their reading, being able to know at every moment their level of understanding. Moreover, when faced to a difficulty, learners can call upon regulation procedures, also called *reading strategies* [1]. Reading strategies have been studied extensively with adolescent and adult readers using the think-aloud procedure that engages the reader to auto-explain at specific breakpoints while reading, therefore providing insight in terms of comprehension.

Four types of reading strategies are mainly used by expert readers [2]. *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).

In addition, teachers need valid and reliable *measures of textual complexity* for selecting texts for the day-to-day instruction. Two approaches compete for the automated of text complexity: 1/ using simple statistical measures that mostly rely on word difficulty (from already-made scales) and sentence length; 2/ using a combination of multiple factors ranging from lexical indicators as word frequency, to syntactic, semantic and even pragmatic levels (e.g., textual cohesion) [3].

As an in-depth perspective, text cohesion, seen as the relatedness between different parts of texts, is a major determinant of text coherence and has been shown to be an important predictor of reading comprehension [4]. Cohesiveness understanding (e.g., referential causal or temporal) is central to the process of building the coherence of a text at the local level, which, in turn, allows the textual content to be reorganized into its macrostructure and situation model at a more global level. High cohesion texts are more beneficial to low-knowledge readers than to high-knowledge readers [5]. Hence, textual cohesion is a feature of textual complexity (through some semantic characteristics of the read text) that might interfere with reading strategies (through the inferences made by a reader).

McNamara and colleagues devised two systems: while *CohMetrix* [5] addresses facets of textual complexity, *iStart* [6] is focused on reading strategies. *CohMetrix* provides a wide range of measures on textual features at five main levels: word (e.g., part-of-speech and frequency), syntax (e.g., percentage of nouns), text-base (e.g., co-

reference and lexical diversity), situation model (e.g., cohesion and temporal indices), and genre and rhetorical structure (e.g., text genre).

iStart is the first implemented system that teaches and assesses self-explanations in accordance to the reading material, with various modules that train learners using the *Self-Explanation Reading Training* method [2]. One module shows how to use those techniques using a virtual student, while another module asks students to read texts and provide verbalizations, evaluates them and gives an appropriate feedback.

ReaderBench encompasses the functionalities of both *CohMetrix* and *iStart*, 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. The main differentiators between *ReaderBench* and previous systems consist of the following: 1/ a generalized model that can be easily extended, in addition to plain essay- or story-like texts, to the analysis of chats and forums, with emphasis on collaboration assessment [7], 2/ different factors, measurements and the use of SVMs for increasing the validity of textual complexity assessment [8], 3/ multi-lingual support and the integration of specific NLP tools for both French and English, and 4/ a different educational purpose, as *ReaderBench* validation was performed on primary school pupils, whereas *iStart* mainly targets high school and university students.

Moreover, the design of *ReaderBench* is focused on 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). 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.

3 Cohesion-based Discourse Analysis

Text cohesion, viewed as lexical, grammatical and semantic overt relationships, is defined within our implemented model in terms of: 1/ the *inverse distance* between textual elements; 2/ *lexical proximity* that is easily identifiable through identical lemmas and semantic distances [9, 10] within ontologies; 3/ semantic similarity measured through *Latent Semantic Analysis* (LSA) [11] and *Latent Dirichlet Allocation* (LDA) [12]. Additionally, specific natural language processing techniques are applied to reduce noise and improve the system's accuracy: tokenizing, splitting, part of speech tagging, parsing, stop words elimination, dictionary-only words selection, stemming, lemmatizing, named entity recognition and co-reference resolution [13].

In order to provide a multi-lingual analysis platform with support for both English and French, *ReaderBench* integrates both *WordNet* [14] and a transposed and serialized version of *WOLF* (*Wordnet Libre du Français*, `http://alpage.inria.fr`

`/~sagot/wolf.html`). 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 [9] and Leacock–Chodorow's normalized path length [10].

Afterwards, semantic models were trained using three specific corpora: "*TextEnfants*" [15] (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 the NLP pipe, the normalization of occurrences through the use of term frequency–inverse document frequency [13] and distributed computing for increasing speedup [16].

LSA and LDA models extract semantic relations from underlying word co-occurrences and are based on the bag of words hypothesis. Although mathematical models behind LSA and LDA are completely different, our experiments have proven that the 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 vector spaces are generated after projecting the matrixes obtained from the reduced-rank Singular Value Decomposition of the initial term-doc matrix and can be used to determine the proximity of words through cosine similarity [11]. From a different viewpoint, LDA topic models provide an inference mechanism of underlying topic structures through a generative probabilistic process [12]. In this context, similarity between concepts can be seen as the opposite of the Jensen-Shannon dissimilarity [13] between their corresponding posterior topic distributions.

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 propose a *cohesion graph* that can be seen as a generalization of the previously proposed utterance graph [17]. More formally, we are building a multi-layered mixed graph consisting of three types of nodes: a central node, the *document* that can represent the entire reading material, *blocks* (paragraphs from the initial text) and *sentences*, the main units of analysis. 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 links* are established between adjacent paragraphs or sentences and are used for best modeling the information flow throughout the discourse, therefore making possible the identification of cohesion gaps. Additional *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.

In contrast, as cohesion can be regarded as the sum of 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, cohesion in itself is not enough to distinguish texts in terms of complexity. However, by shifting the sense of inter-dependencies, a facet of textual complexity is strongly tied to cohesion. In order to better highlight this perspective, 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.

As a *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 standard deviation between raters was of .80. In the end, 540 individual cohesion scores were collected and were 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).

The previous results show that the proposed combined method of integrating multiple semantic similarity measures outperforms all individual metrics, that a greater 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 approaches and the complementarity effect in terms of the reduction of errors that can be induced by using a single method.

4 Reading Strategies

Starting from the four main types of reading strategies introduced in section 2, our aim was to integrate automatic extraction methods designed to support tutors at identifying various strategies and to best fit with the annotation methodology aligned with [2]. We have tested various methods of identifying reading strategies and we will focus solely on presenting the alternatives that provided the best overall correlations.

In ascending order of complexity, the simplest strategies to identify are *causality* and *control* 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, rather than creating an inferential link. Afterwards, *paraphrases*, that were considered repetitions of the same lexical structures by human raters, were automatically identified based on word lemmas and synonymy relationships from the lexicalized ontologies.

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 text. The latter distances had to be taken into consideration for better weighting the importance of each concept, with respect to the whole text.

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 reference reading material. If more than 2 similarity measures were above the mean value and exceeded a minimum threshold, bridging was estimated as the number of links between contiguous zones of cohesive sentences.

Figure 1 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.

We ran an experiment with pupils aged from 9 to 11 (grades 3-5) who had to read a 450 word-long story 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 evaluated by two human experts, and categorized according to McNamara [2]'s scoring scheme. 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 [18].

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. Moreover, "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 solely looking at manual assessments, high discrepancies between evaluators were identified due to different understandings and perceptions of pupil's intentions expressed within their meta-cognitions. 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.

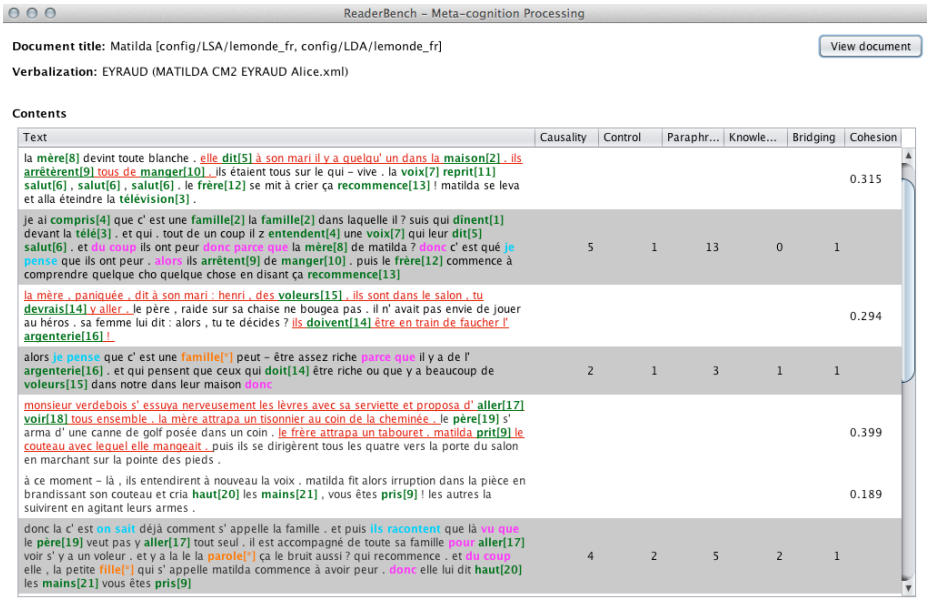

Figure 1. Reading strategies analysis in *ReaderBench.*

5 Textual Complexity

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). 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. We propose a multi-dimensional analysis of textual complexity, covering a multitude of factors depicted in Table 2 (extensive description in [8]) aggregated through the use of Support Vector Machines, which has proven to be the most efficient [19].

Hence, besides the factors presented in [8] that were focused on a more shallow approach, of particular interest is how semantic and pragmatic factors correlate to classic readability measures. Therefore, starting from the textual complexity model that already integrated classic readability formulas, surface metrics derived from classic automatic essay grading techniques, morphology and syntax factors [8], we have introduced new classes focused on semantics and pragmatics. Firstly, *cohesion* reflected in the strength of inner-block and inter-block links influences readability, as semantic similarities govern the understanding of a text. Secondly, a variety of metrics based on the span and the coverage of *lexical chains* [20] provide insight in

terms of lexicon variety and of cohesion, expressed in this context as the semantic distance between different chains. Thirdly, *entity-density features* proved to influence readability as the number of entities introduced within a text is correlated to the working memory on the text's targeted readers. Finally, another dimension focuses on the ability to resolve referential relations correctly [21] as *co-reference inference features* also impact comprehension difficulty (e.g., the overall number of chains, the inference distance or the span between concepts). 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 no corpus was available for French in order to train our complexity model, we have opted to automatically extract texts from TASA, using its Degree of Reading Power (DRP) score, into six classes of complexity [22]. 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 features* at semantic and pragmatic levels *provide relevant insight* that can be used for automatic classification. As particular implementation aspects for increasing the effectiveness of SVMs, all factors were linearly scaled and a Grid Search optimization method was enforced. In the end, *k*-fold cross validation [23] was applied for extracting the following performance features (see Table 1): *precision* or *exact agreement* (EA) and *adjacent agreement* (AA), as the percent to which the SVM was close to predicting the correct classification.

Table 1. Textual complexity classes.

Depth of metrics	Classes of factors for evaluation	Avg. *EA*	Avg. *AA*
Surface Analysis	Readability formulas	.717	.995
	Fluency factors	.314	.579
	Structure complexity factors	.728	.993
	Diction factors	.550	.901
	Entropy factors (words vs. characters)	.313	.573
	Word complexity factors	.556	.918
Morphology & Syntax	Balanced CAF (Complexity, Accuracy, Fluency)	.755	.996
	Specific POS complexity factors	.570	.929
	Parsing tree complexity factors	.424	.806
Semantics & Pragmatics	Cohesion through lexical chains, LSA and LDA	.544	.894
	Named entity complexity factors	.590	.929
	Co-reference complexity factors	.384	.730
	Lexical chains	.367	.704

Moreover, two additional measurements were performed. Firstly, an integration of all metrics from all complexity classes proved that the SVMs results are compatible with the DRP scores (*EA* = .763 and *AA* = .997), and that they provide significant im-

provements as they outperform any individual class precisions. The second measurement (*EA* = .597 and *AA* = .943) uses solely morphology, semantics and pragmatics measures in order to avoid a circular comparison, as the DRP score is based on shallow factors. This result shows a link between low-level factors (also used in the DRP score) and in-depth analysis factors, that can also be used to accurately predict the complexity of a reading material.

6 Conclusion and Future Research Directions

ReaderBench is an environment integrating new ways to assess a wide range of cognitive processes involved in reading through the use of advanced NLP techniques. It provides a semantic insight and discourse structure through the combination of multiple semantic distances. Its flexibility and extensibility make it easily integrable in various educational settings to foster self-regulated learning. With further improvements, like chat/forum collaboration assessment, a human-rated corpus for textual complexity SVM training, as well as a speech-to-text functionality enabling its use with younger pupils, *ReaderBench* will effectively support students in their learning and CSCL activities.

7 Acknowledgements

This research was supported by an *Agence Nationale de la Recherche* (DEVCOMP) grant, by the 264207 ERRIC–Empowering Romanian Research on Intelligent Information Technologies/FP7-REGPOT-2010-1 and the POSDRU/107/1.5/S/76909 Harnessing human capital in research through doctoral scholarships (ValueDoc) projects.

8 References

1. McNamara, D.S. and Magliano, J.P.: Self-explanation and metacognition. In: Hacher, J.D., Dunlosky, J. and Graesser, A.C. (eds.) *Handbook of metacognition in education*, pp. 60–81. Erlbaum, Mahwah (2009)
2. McNamara, D.: SERT: Self-Explanation Reading Training. *Discourse Proc.*, 38, 1–30 (2004)
3. Nelson, J., Perfetti, C., Liben, D. and Liben, M.: Measures of text difficulty. Technical Report to the Gates Foundation (2011)
4. Tapiero, I.: *Situation models and levels of coherence*. Erlbaum, Mahwah (2007)
5. McNamara, D.S., Louwerse, M.M., McCarthy, P.M. and Graesser, A.C.: Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Proc.*, 47(4), 292–330 (2010)
6. McNamara, D., Boonthum, C. and Levinstein, I.: Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In: Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (eds.) *Handbook of Latent Semantic Analysis*, pp. 227–241. Erlbaum, Mahwah (2007)

7. Trausan-Matu, S., Dascalu, M. and Rebedea, T.: A system for automatic analysis of Computer-Supported Collaborative Learning chats. In: *12th Conf. ICALT*, pp. 95–99. IEEE (2012)
8. Dascalu, M., Trausan-Matu, S. and Dessus, P.: Towards an integrated approach for evaluating textual complexity for learning purposes. In: Popescu, E., Klamma, R., Leung, H. and Specht, M. (eds.) *Advances in web-based learning (ICWL 2012)*, Vol. LNCS 7558, pp. 268–278. Springer, New York (2012)
9. Wu, Z. and Palmer, M.: Verb semantics and lexical selection. In: *32nd Annual Meeting of the Association for Computational Linguistics*, pp. 133–138. ACL, Las Cruces (1994)
10. Leacock, C. and Chodorow, M.: Combining local context and WordNet similarity for wordsense identification. In: Fellbaum, C. (ed.) *WordNet: An electronic lexical database*, pp. 265–283. MIT Press, Cambridge (1998)
11. Landauer, T.K. and Dumais, S.T.: A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychol. Rev.*, 104(2), 211–240 (1997)
12. Blei, D., Ng, A. and Jordan, M.: Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5), 993–1022 (2003)
13. Manning, C. and Schütze, H.: *Foundations of statistical Natural Language Processing*. MIT Press, Cambridge (Mass.) (1999)
14. Miller, G.A.: *WordNet*: A Lexical Database for English. *Comm. ACM*, 38(11), 39–41 (1995)
15. Denhière, G., Lemaire, B., Bellissens, C. and Jhean-Larose, S.: A semantic space for modeling children's semantic memory. In: Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (eds.) *Handbook of Latent Semantic Analysis*, pp. 143–165. Erlbaum, Mahwah (2007)
16. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit, http://mallet.cs.umass.edu (2002)
17. Trausan-Matu, S., Dascalu, M. and Dessus, P.: Considering textual complexity and comprehension in Computer-Supported Collaborative Learning. In: Cerri, S.A., Clancey, W.J., Papadourakis, G. and Panourgia, K. (eds.) *11th Int. Conf. on Intelligent Tutoring Systems (ITS 2012)*, Vol. LNCS 7315, pp. 352–357. Springer, New York (2012)
18. Graesser, A.C., Singer, M. and Trabasso, T.: Constructing inferences during narrative text comprehension. *Psychol. Rev.*, 101(3), 371–395 (1994)
19. François, T. and Miltsakaki, E.: Do NLP and machine learning improve traditional readability formulas? In: *Proc. First Workshop on Predicting and improving text readability for target reader populations (PITR2012)*, pp. 49-57. ACL, Montréal (2012)
20. Galley, M. and McKeown, K.: Improving word sense disambiguation in lexical chaining. In: *18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, Acapulco (2003)
21. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M. and Jurafsky, D.: Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: *15th Conference on Computational Natural Language Learning*, pp. 28–34 (2011)
22. McNamara, D.S., Graesser, A.C. and Louwerse, M.M.: Sources of text difficulty: Across the ages and genres. In: Sabatini, J.P. and Albro, E. (eds.) *Assessing reading in the 21st century*. R&L Education, Lanham (in press)
23. Geisser, S.: *Predictive Inference*. Chapman and Hall, New York (1993)