



Mining texts, learners productions and strategies with ReaderBench

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

► To cite this version:

Mihai Dascalu, Philippe Dessus, Maryse Bianco, Stefan Trausan-Matu, Aurélie Nardy. Mining texts, learners productions and strategies with ReaderBench. Alejandro Pena-Ayala. Educational Data Mining: Applications and Trends, Springer, pp.345-377, 2014, 10.1007/978-3-319-02738-8_13. hal-00979702

HAL Id: hal-00979702

<https://hal.science/hal-00979702>

Submitted on 22 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mining Texts, Learners Productions and Strategies with *ReaderBench*

Mihai Dascălu¹, Philippe Dessus², Maryse Bianco²,
Ștefan Trăușan-Matu¹ and Aurélie Nardy²

¹ Politehnica University of Bucharest, Computer Science Department, Romania
{mihai.dascalu, stefan.trausan}@cs.pub.ro

² Univ. Grenoble Alpes, LSE, France
{philippe.dessus, maryse.bianco, aurelie.nardy}@upmf-grenoble.fr

Abstract. The chapter presents *ReaderBench*, a multi-lingual and flexible environment that integrates text mining technologies for assessing a wide range of learners' productions and for supporting teachers in several ways. *ReaderBench* offers three main functionalities in analyzing texts: cohesion-based assessment, reading strategies identification, and textual complexity evaluation. All of these have been object to empirical validations. *ReaderBench* may be used during an entire educational scenario, starting from the initial complexity assessment of the reading materials, the assignment of texts to learners, the detection of reading strategies reflected in one's self-explanations, and comprehension evaluation fostering learner's self-regulation process.

1 Introduction

Recent NLP techniques, as well as the ever-growing computer power, enable the design and implementation of new systems that automatically deliver summative and formative assessments to learners, using multiple sets of data (e.g., textual material, behavior tracks, meta-cognitive explanations). New automatic evaluation processes allow teachers and learners to have immediate information on how they learn or understand. Furthermore, computer-based systems can be integrated into pedagogical scenarios providing activity flows that foster learning.

ReaderBench is a fully functional framework based on text mining technologies [1] that can be seen as a cohesion-based integrated approach which addresses multiple dimensions of learner comprehension, including the identification of reading strategies, textual complexity assessment and even CSCL, with emphasis on participant involvement and collaboration [2] – this latter facility will not be introduced in this chapter for readability's sake. *ReaderBench* provides teachers and learners information on their reading/writing activities: initial textual

complexity assessment, assignment of texts to learners, capture of self-explanations reflected in pupil's textual verbalizations, and reading strategies assessment [2].

The remainder of this chapter is as follows. The next section introduces a general perspective over data and text mining approaches used in educational applications. The third section is an overview on how learner comprehension can be modeled and predicted. The fourth section presents the core text feature from which almost all *ReaderBench* measures are computed: cohesion. The next four sections present the main functionalities of our system: topic extraction, cohesion analysis, reading strategies analysis and textual complexity assessment. The ninth section introduces an educational scenario that gives substance and real-world use of *ReaderBench*.

2 Data and Text Mining for Educational Applications

Learning analytics aims at measuring, collecting, analyzing and “reporting data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occur” (Society for Learning Analytics Research, <http://www.solaresearch.org/>). While the main focus of this approach is to gather data about learners (e.g., their behavior, opinions, affects, social interactions), very few research is performed to infer what learners actually understand and the learning contexts are rarely taken into account (e.g., which learning material is used, to what pedagogical intent, within which educational scenario) [3].

Educational data analyzed from computer-based approaches typically comes from two wide categories: *knowledge* (e.g., textual material from course information) and *behavior* (e.g., learners' behavior in LMSs from log analysis). Whereas a substantial amount of research is centered on behavioral data [4], relatively few research encompasses the analysis of textual materials, presented beforehand to learners. Raw data is ideally and easily computable with data mining techniques [5], but inferences from these data to uncover learners' cognitive processes are far more complex and involve comparisons to human experts judgments.

Our approach stems from a very broad idea. 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 [6]. In turn, cohesion analyses can be applied in a wide range of data analyses in educational contexts: text readability and difficulty, knowledge relatedness, chat or forum group replies. The next section addresses learner comprehension, its relationships with textual complexity, and how it can be inferred from learner's self-explanations.

3 Predicting Learner Comprehension

Learner's comprehension of textual materials during reading depends both on text properties and on the learner's reading skills. It has long been recognized that the comprehension performance differs according to lexical and syntactical complexity, as well as to the thematic content and to how information is structured [7, 8]. Of particular importance are the cohesion and coherence properties of texts that can help or impair [9] and, moreover, interact with reader's personal characteristics [8, 10]. On the reader's side, her background knowledge and the reading strategies she is able to use to process information are also strong predictors of reading comprehension, in addition to her word recognition ability and semantic skills [11, 12, 13]. The remainder of this section elaborates more on the two main factors of text understanding: textual features (through textual complexity), and readers' abilities (through reading strategies).

3.1 Textual Complexity Assessment for Comprehension Prediction

Teachers usually need valid and reliable measures of textual complexity for their day-to-day instruction in order to select materials appropriate to learners' reading level. This proves to be a challenging and cumbersome activity since different types of texts (narrative, argumentative or expository) place different demands on different reading skills [14, 15]. For example, McNamara and her colleagues [15] found that narrative texts contain more familiar words than scientific texts, but that they have more complex syntactic sentences, as well. Narratives were also found to be less cohesive than science expository texts, the latter more strongly requiring background knowledge. In conclusion, different skills must be involved in comprehending different types of text and the same reader can be more or less able to comprehend a text corresponding to her reading and/or grade level.

Two approaches usually compete for the automated assessment 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 and semantic levels (e.g., textual cohesion) [16].

As an in-depth perspective, text cohesion, seen as the relatedness between different parts of texts, is a major determinant for building a coherent representation of discourse and has been shown to be an important predictor of reading comprehension [6]. Cohesiveness understanding (e.g., referential, causal or temporal) is central to the process of building textual coherence at local level, which, in turn, allows the textual content to be reorganized into its macrostructure and situation model at global level. Highly cohesive texts are more beneficial to low-knowledge readers than to high-knowledge ones [17]. 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). Moreover, inference skills and the ability to plan and organize

information have been shown to be strongly tied to the comprehension performance of more complex texts [14]. These findings let us consider cohesion as one of the core determinants of textual complexity.

3.2 The Impact of Reading Strategies extracted from Self-Explanations for Comprehension Assessment

Expert readers are strategic readers. They monitor their reading, being able to know at every moment their level of understanding. When faced with a difficulty, learners can call upon regulation procedures, also called reading strategies [18]. 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 the comprehension mechanisms they call upon to interpret the information they are reading. In other words, reading strategies are defined here as “the mental processes that are implicated during reading, as the reader attempts to make sense of the printed words” [19, p. 40].

Four types of reading strategies are mainly used by expert readers [20]. *Paraphrasing* allows the reader to express what 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 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).

We recently performed an experiment [21] to extend the assessment of reading strategies with children ranging from 3rd to 5th grade (8–11 years old). Children 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 McNamara’s [20] coding scheme, covering the following strategy items: paraphrases, textual inferences, knowledge inferences, self-evaluations, and “other”. The “other” category is very close to the “irrelevant” category [20] as it aggregates irrelevant, as well as unintelligible statements. Two dominant strategies were identified: paraphrases and text-based inferences; text-based inferences frequency increases from grade 3 to 5, while erroneous paraphrases frequency decreases; knowledge-based inferences remain rare, but their frequency doubled from grade 3 to 5, amounting from 4 to 8% of the identified reading strategies within the appropriate verbalizations.

Three results are noteworthy. Firstly, self-explanations are a useful tool to access the reading strategies of young children (8–11 years old) who already

dispose of all the strategies older children carry out. Secondly, we found a relation 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. A better comprehension in this age range is tied to less false paraphrases and more text-based inferences ($R^2 = .18$ for paraphrases and $R^2 = .16$ for text-based inferences). Thirdly, mediation models [22] showed that verbal ability partially mediates the effect of text-based inferences and that age moderates this mediating effect. The effect of text-based inferences on reading comprehension is mediated by verbal ability for the younger students while it becomes a direct effect for older students.

Starting from the previous experiments and literature findings, one of the goals of *ReaderBench* is to enable the usage of new texts with little or no human intervention, providing both textual complexity assessments on these texts, and a fully automatic identification of reading strategies as a support for teachers. The textual complexity assessment aims at calibrating texts before providing them to learners.

4 Cohesion-Based Discourse Analysis

4.1 Construction of the Coherence Graph

Text cohesion, viewed as lexical, grammatical and semantic relationships that link together textual units, is defined within our implemented model in terms of: 1/ the *inverse normalized distance between textual elements* expressed in terms of the number of textual analysis elements in-between; 2/ *lexical proximity* that is easily identifiable through identical lemmas and semantic distances within ontologies [23]; 3/ *semantic similarity* measured through Latent Semantic Analysis (LSA) [24] and Latent Dirichlet Allocation (LDA) [25]. Additionally, specific natural language processing techniques [26] are applied to reduce noise and improve the system’s accuracy: spell-checking (optional) [27, 28], tokenizing, splitting, part of speech tagging [29, 30], parsing [31, 32], stop words elimination, dictionary-only words selection, stemming [33], lemmatizing [34], named entity recognition [35] and co-reference resolution [36, 37].

In order to provide a multi-lingual analysis platform with support for both English and French, *ReaderBench* integrates both *WordNet* [38] and a transposed and serialized version of *Wordnet Libre du Français (WOLF)* [39]. 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 [40] and Leacock–Chodorow’s normalized path length [41].

Afterwards, LSA and LDA semantic models were trained using three specific corpora: “*TextEnfants*” [42] (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) [43, 44, 45], the normalization of occurrences through the use of term frequency-inverse document frequency (*Tf-Idf*) [26] and distributed computing for increasing speedup [46, 47].

LSA and LDA models extract semantic closeness relations from underlying word co-occurrences and are based on the bag-of-words hypothesis. 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 [24]. From a different viewpoint, LDA topic models provide an inference mechanism of underlying topic structures through a generative probabilistic process [25]. In this context, similarity between concepts can be seen as the opposite of the Jensen-Shannon dissimilarity [26] between their corresponding posterior topic distributions.

From a computational perspective, the LSA semantic spaces were trained using a Tagged LSA engine [43] that preprocesses all training corpora (stop-words elimination, POS tagging, lemmatization) [44, 45], applies *Tf-Idf* and uses a distributed architecture [46, 48] to perform the Singular Values Decomposition. With regards to LDA, the parallel topics model used iterative Gibbs sampling over the training corpora [47] with 10,000 iterations and 100 topics, as recommended by [25]. 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 [2, 49] that can be seen as a generalization of the previously proposed utterance graph [50, 51, 52]. More formally, we are building a multi-layered mixed graph consisting of three types of nodes [53]: 1/ a central node, the *document* that represents the entire reading material, 2/ *blocks*, a generic entity that can reflect paragraphs from the initial text and 3/ *sentences*, the main units of analysis, seen as collections of words and grammatical structures obtained after the initial NLP processing.

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*. *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.

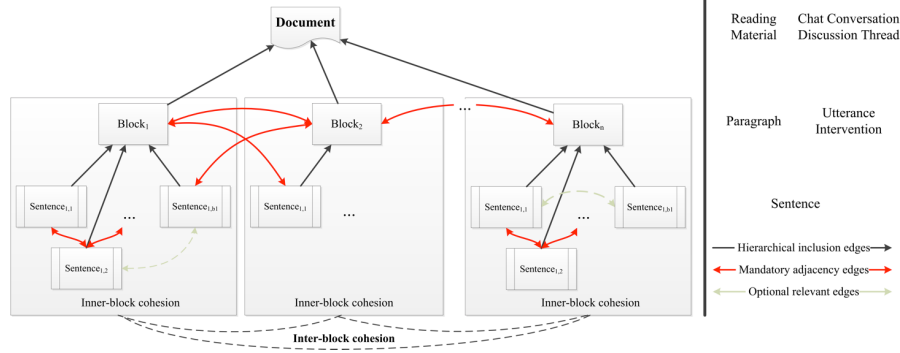


Fig. 1. The cohesion graph as underlying discourse structure.

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 [54].

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.

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, 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.

4.2 Validation of the Cohesion Measure

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.

5 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, 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:

1. *Individual normalized term frequency* $-I + \log(\text{no_occurrences})$ [55]; 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.
2. *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.
3. 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* [56] 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 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. 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, **Fig. 2** depicts an excerpt from [57] presented to 1st year master students during the Natural Language Processing Course 2011-2012.

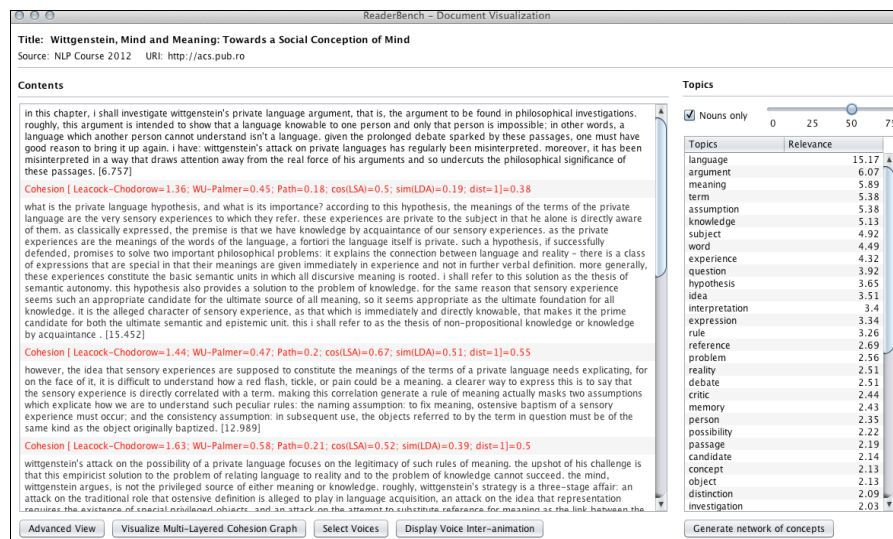


Fig. 2. *ReaderBench* main interface for visualizing documents and topics.

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* [58] and the dimension of each concept is proportional with its betweenness score [59] from the generated network.

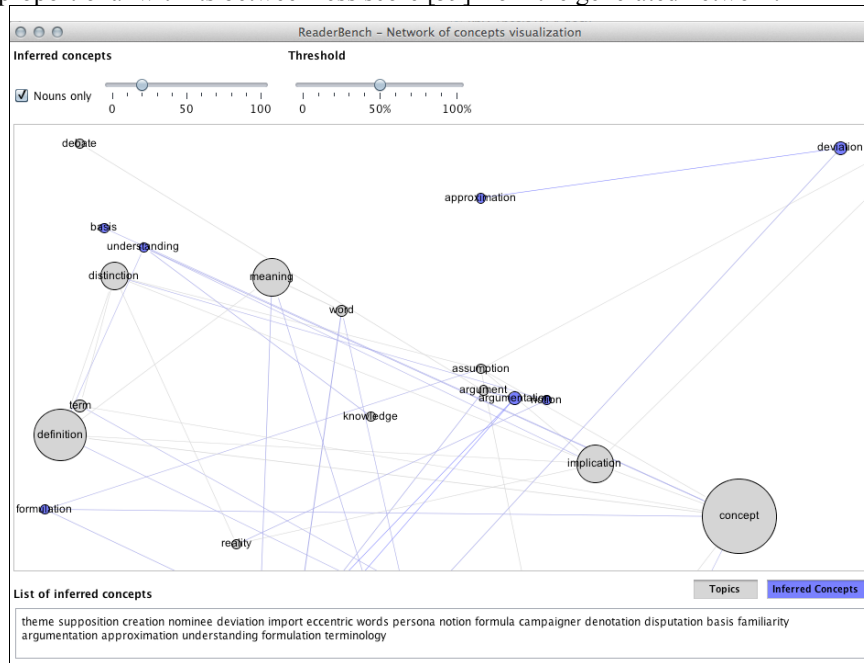


Fig. 3. Network of concepts visualization from and inferred from [57].

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 **Fig. 3**, “*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. 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.

6 Cohesion-Based Scoring Mechanism of the Analysis Elements

A central component in the evaluation process of each sentence's importance is our bottom-up scoring method. Although tightly related to the cohesion graph [60] that is browsed from bottom to top and is used for augmenting the importance of the analysis elements, the initial assessment of each element is based on its topics coverage and their corresponding relevance, with respects to the entire document. 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 presented in the previous section. 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. 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.

Fig. 4 presents the main user interface of *ReaderBench* highlighting the following elements: block scores (in square brackets after each paragraph), demarcation with bold of sentences considered most important according to the summarization facility and document topics and identified topics ordered by relevance. 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.

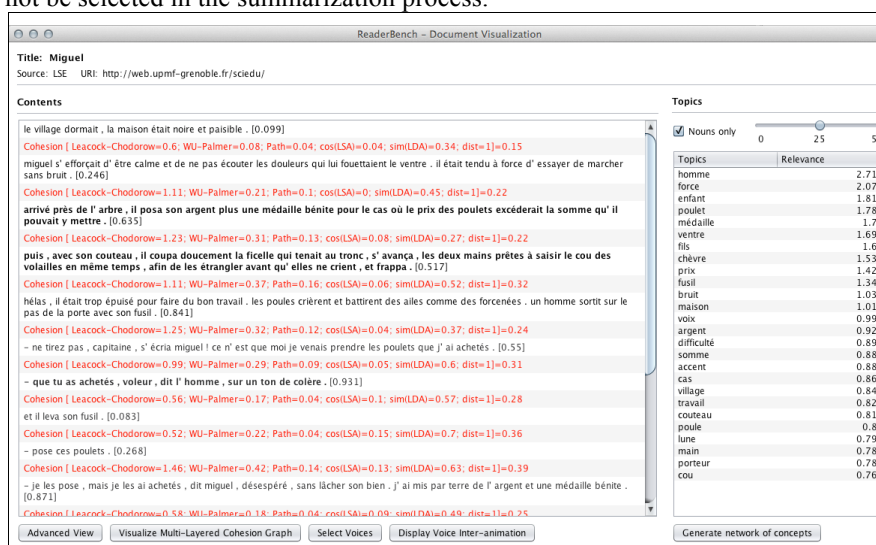


Fig. 4. Reading material visualization.

In addition, starting from tutors' general observations that an extractive summarization facility, combined with the demarcation of the most important sentences, is useful for providing a quick overview of the reading material, 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* [61] 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 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.

As preliminary validation we have performed experiments on two narrative texts in French: “*Miguel de la faim*” [62] and “*La pharmacie des éléphants*” [63], starting from the measurements initially performed by [64] in which 330 high school (9th–12th grade) students and 25 tutors were asked to manually highlight the most important 3 to 5 sentences from the two presented stories [64]. The inter-rater agreement scores were rather low, as the ICC (Intraclass Correlation Coefficient) values were of .13, respectively .23, highlighting the subjectivity of the task at hand.

Afterwards, as suggested by [65], four equivalence classes were defined, taking into consideration the mean – standard deviation, mean and mean + standard deviation 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 1**) 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 1. Exact and Adjacent Agreement between automatic and manual sentence selection using equivalence classes.

Text	Exact/Adjacent Agreement (EA/AA)					Avg. EA/ AA
	9th grade	10th grade	11th grade	12th grade	Tutor	
<i>Miguel de la faim</i>	.33/.83	.42/.75	.29/.88	.38/.88	.46/.88	.38/.84
<i>La pharmacie des éléphants</i>	.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 Reading Strategies Identification Heuristics

7.1 Heuristics used to identify Control, Causality, Paraphrasing, Bridging and Knowledge Inference

Starting from the two previous studies and the five types of reading strategies used by [66], 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 two observed differences. Firstly, very few predictions were used, perhaps due to the age of the pupils, compared to McNamara’s subjects; secondly, 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 [29].

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.

ReaderBench – Meta-cognition Processing

Document title: Matilda [config/LSA/lemonde_fr, config/LDA/lemonde_fr]

Verbalization: (MATILDA CM2) (lemonde_fr.xml)

View document

Contents

Text	Causality	Control	Paraphr...	Knowle...	Bridging	Cohesion
la mère[8] devint toute blanche. elle dit[5] à son mari il y a quelqu'un dans la maison[2]. ils arrêtèrent[9] tous de manger[10]. ils étaient tous sur le qui-vive. la voix[7] reprit[11]. salut[6]. salut[6]. salut[6]. le frère[12] se mit à crier ça recommence[13] ! matilda se leva et alla éteindre la télévision[3].						0.315
je ai compris[4] que c'est une famille[2] la famille[2] dans laquelle il ? suis qui dinent[1] devant la télé[3]. et qui. tout de un coup il z entendent[4] une voix[7] qui leur dit[5] salut[6]. et ils ont peur parce que la mère[8] de matilda ? pense c'est qu'il y a quelqu'un qui les ont peur. ils arrêtèrent[9] de manger[10]. puis le frère[12] commence à comprendre quelque chose en disant ça recommence[13].	5	1	13	0	1	0.294
la mère. paniquée. dit à son mari. henri. des voleurs[15]. ils sont dans le salon. tu devrais[14] 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[14] être en train de faucher l'argenterie[16].						0.399
alors je pense que c'est une famille[2] peut-être assez riche parce que il y a de l'argenterie[16]. et qui pensent que ceux qui doit[14] être riche ou que y a beaucoup de voleurs[15] dans notre dans leur maison d'ailleurs.	2	1	3	1	1	0.189
monsieur verdebois s'essuya nerveusement les lèvres avec sa serviette et proposa d'aller[17] voir[18] tous ensemble. la mère attrapa un tisonnier au coin de la cheminée. le père[19] s'arma d'une canne de golf posée dans un coin. le frère attrapa un tabouret. matilda prit[9] 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[20] les mains[21], vous êtes pris[9] ! les autres la suivirent en agitant leurs armes.						
donc là c'est un fait déjà comment s'appelle la famille. et puis ils racontent que là le père[19] veut pas y aller[17] tout seul. il est accompagné de toute sa famille. aller[17] voir s'y a un voleur. et y a la le la parolot[1] ça le bruit aussi ? qui recommence. et ils disent elle. la petite fille[1] qui s'appelle matilda commence à avoir peur. donc elle lui dit haut[20] les mains[21] vous êtes pris[9].	4	2	5	2	1	

Fig. 5. Visualization of automatically identified reading strategies.

Fig. 5 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 grey sections represent the pupil's self-explanations, whereas the white blocks represent paragraphs from "Matilda" [67]. **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.

7.2 Strategies Identification Validation

We ran an experiment with pupils aged from 9 to 11 who had to read aloud a 450 word-long story, "Matilda" [67], 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 [68].

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.

8 Textual Complexity Assessment

8.1 Multi-Dimensional Integrated Model for Assessing 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.

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 [69], as well as new dimensions focused on semantics [60]. In the end, subsets of specific factors are aggregated through the use of Support Vector Machines [70], which has proven to be the most efficient [71, 72]. In order to provide an overview, the textual complexity dimensions, with their corresponding performance scores, are presented in **Table 2**, whereas the following paragraphs focus solely on the semantic dimension of the analysis. In other words, besides the factors presented in detail in [69] that were focused on a more shallow approach, of particular interest is how semantic factors correlate to classic readability measures [60].

Table 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
	Balanced CAF (Complexity, Accuracy, Fluency)	.752	.997
Morphology & Syntax	Specific POS complexity factors	.563	.931
	Parsing tree complexity factors	.416	.792
	Cohesion through lexical chains, LSA and LDA	.526	.891
Semantics	Named entity complexity factors	.575	.922
	Co-reference complexity factors	.366	.738
	Lexical chains	.363	.714

Firstly, *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 [73]. 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 [2, 60].

Secondly, a variety of metrics based on the *span* and the *coverage of lexical chains* [56] 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 [74]. 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 [36, 75] 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 Validation of Textual Complexity Assessment

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 [15] of equal frequency, as no corpus was available for French (see **Table 3**).

Table 3. Ranges of the DRP scores as a function of defining the six textual complexity classes [after 15].

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 [76] was applied for extracting the following performance features (see **Table 2**): precision or exact agreement (EA) and adjacent agreement (AA) [71], as the percent to which the SVM was close to predicting the correct classification.

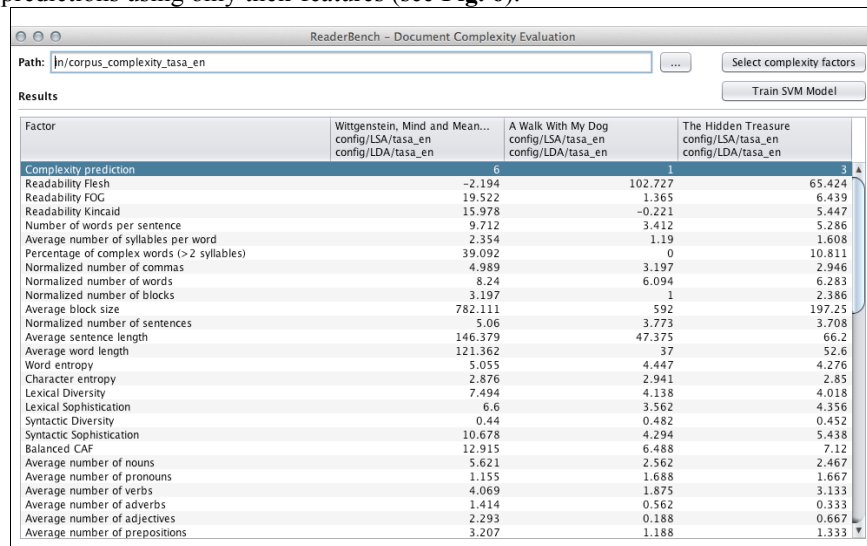
By considering the granular factors, although simple in nature, readability formulas, the average number of words per sentence, the average length of sentences/words and balanced CAF provided the best alternatives at lexical and syntactic level; this was expected as the DRP score is based solely on shallow 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 do not 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.

Two additional measurements were performed in the end. 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.

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. Moreover, 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 **Fig. 6**).



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

Fig. 6. Document complexity evaluation.

9 An Educational Scenario

ReaderBench can be used in a wide range of educational situations and plays the role of a Personal Learning Environment (PLE), allowing three kinds of work-loops, in which teacher/learners can be freely involved, thus triggering self-regulated activities [77]. It is worth noting that these three loops do not generate behavioral data per se, to be analyzed in turn in the software. The first loop is related to texts, the second and the third to both learners' productions and strategies (see **Fig. 7**).

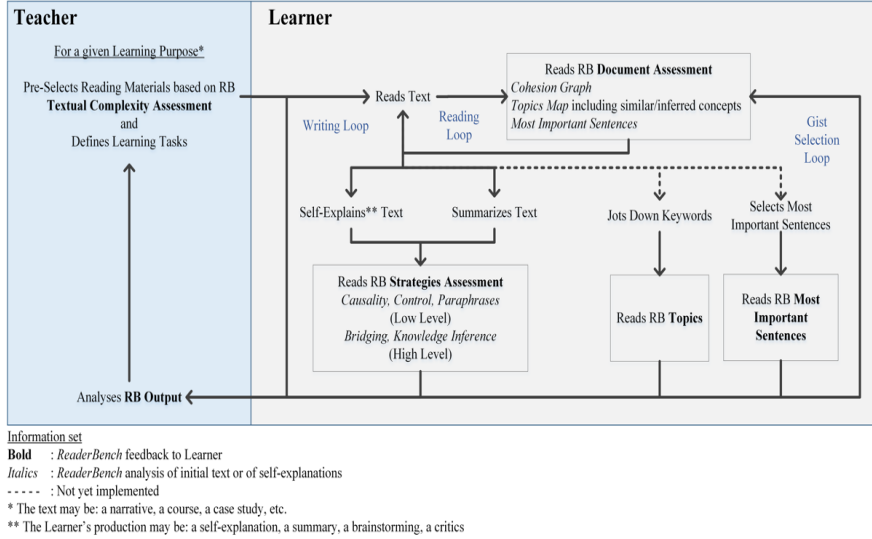


Fig. 7. Learner centered educational scenario in *ReaderBench*.

The first one is a *reading loop*: learners read some material (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*, which is a bit more interactive than the previous. 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*, which gives learners opportunity to develop at length what they understood from the text (e.g., summaries) or the way they understood (strategies self-explanation). Besides these three loops, the teacher can use *ReaderBench* to select appropriate textual materials according to learners' level.

10 Conclusion

We introduced *ReaderBench*, a multi-lingual and multi-purpose system which allows learners and teachers to mine and analyze textual materials, learners' productions and identify reading strategies. This system allows a large range of measures that have been carefully compared to human ones. Moreover, it infers cognitive processes engaged in understanding and can be integrated in several pedagogical scenarios.

Further research will lead to the use of *ReaderBench* in classrooms by teachers and learners in order to validate the pedagogical scenarios. Moreover, the large range of raw data generated by *ReaderBench* will be subject to analysis in an educational data mining platform, like *UnderTracks* [78].

11 Acknowledgements

This research was supported by an Agence Nationale de la Recherche (ANR-10-BLAN-1907) 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. We also wish to thank Sonia Mandin, who kindly provided experimental data used for the validation of sentence importance. Some parts of this paper stem from [53].

12 References

1. Agrawal, R., Batra, M.: A detailed study on text mining techniques. *International Journal of Soft Computing and Engineering*, 2(6), 118–121 (2013)
2. Trausan-Matu, S., Dascalu, M., Dessus, P.: Textual Complexity and Discourse Structure in Computer-Supported Collaborative Learning. In: 11th Int. Conf. on Intelligent Tutoring Systems (ITS 2012), Vol. LNCS 7315, pp. 352–357. Springer, Chania, Grece (2012)
3. Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A data repository for the EDM community: The PSLC datashop. In: Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (eds.) *Handbook of Educational Data Mining*, pp. 43–55. Taylor & Francis, Boca Raton (2011)
4. Zou, M., Xu, Y., Nesbit, J.C., Winne, P.H.: Sequential pattern analysis of learning logs: Methodology and applications. In: Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (eds.) *Handbook of Educational Data Mining*, pp. 107–121. Taylor & Francis, Boca Raton (2011)
5. Sheard, J.: A data repository for the EDM community: The PSLC datashop. In: Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (eds.) *Handbook of Educational Data Mining*, pp. 27–42. Taylor & Francis, Boca Raton (2011)
6. Tapiero, I.: Situation models and levels of coherence. Erlbaum, Mahwah, NJ (2007)
7. Schnotz, W.: Comparative Instructional text organization. In: Mandl, H., Stein, N.L., Trabasso, T. (eds.) *Learning and comprehension of text* pp. 53–81. Lawrence Erlbaum, Hillsdale (1984)
8. McNamara, D., Kintsch, E., Songer, N.B., Kintsch, W.: 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 (1996)
9. Oakhill, J., Garnham, A.: On theories of belief bias in syllogistic reasoning. *Cognition*, 46(1), 87–92 (1993)
10. O'Reilly, T., McNamara, D.S.: Reversing the reverse cohesion effect: good texts can be better for strategic, high-knowledge readers. *Disourse Processes*, 43(2), 121–152 (2007)
11. Cain, K., Oakhill, J.: Reading comprehension development from 8 to 14 years, the contribution of component skills and processes. In: Wagner, R.K., Schatschneider, C.,

- Phythian-Sence, C. (eds.) *Beyond decoding, the behavioral and biological foundations of reading comprehension*, pp. 143–175. Guilford Press, New York (2009)
12. Kintsch, W.: *Comprehension, a paradigm for cognition*. Cambridge University Press, Cambridge (1998)
 13. McNamara, D.S., O'Reilly, T.: Theories of comprehension skill: knowledge and strategies versus capacity and suppression. In: Colombus, A.M. (ed.) *Progress in experimental psychology research*, pp. 113–136. Nova Science Publishers, Hauppauge (2009)
 14. Eason, S.H., Goldberg, L., Cutting, L.: Reader-text interactions: how differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology*, 104(3), 515–528 (2012)
 15. McNamara, D.S., Graesser, A.C., Louwerse, M.M.: Sources of text difficulty: Across the ages and genres. In: Sabatini, J.P., Albro, E. (eds.) *Assessing reading in the 21st century*. R&L Education, Lanham (in press)
 16. Nelson, J., Perfetti, C., Liben, D., Liben, M.: Measures of text difficulty. Technical Report to the Gates Foundation (2011)
 17. McNamara, D.S., Louwerse, M.M., McCarthy, P.M., Graesser, A.C.: Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Proc.*, 47(4), 292–330 (2010)
 18. McNamara, D.S., Magliano, J.P.: Self-explanation and metacognition. In: Hacher, J.D., Dunlosky, J., Graesser, A.C. (eds.) *Handbook of metacognition in education*, pp. 60–81. Erlbaum, Mahwah (2009)
 19. Millis, K., Magliano, J.: Assessing comprehension processes during reading. In: Sabatini, J.P., O'Reilly, T., Albro, E.R. (eds.) *Reaching an understanding*, pp. 35–54. Rowman & Littlefield, Lanham (2012)
 20. McNamara, D.S.: SERT: Self-Explanation Reading Training. *Discourse Processes*, 38, 1–30 (2004)
 21. Nardy, A., Bianco, M., Toffa, F., Rémond, M., Dessus, P.: Contrôle et régulation de la compréhension : L'acquisition de stratégies de 8 à 11 ans. In: David, J., Royer, C. (eds.) *L'apprentissage de la lecture : convergences, innovations, perspectives*. Peter Lang, Berne (in press)
 22. Hayes, A.F. (ed.): *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* The Guilford Press, New York, NY (2013)
 23. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1), 13–47 (2006)
 24. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2), 211–240 (1997)
 25. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5), 993–1022 (2003)
 26. Manning, C.D., Schütze, H.: *Foundations of statistical Natural Language Processing*. MIT Press, Cambridge, MA (1999)
 27. Alias-i: LingPipe, <http://alias-i.com/lingpipe> (2008)
 28. McCandless, M., Hatcher, E., Gospodnetic, O.: *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, USA (2010)

29. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: HLT-NAACL 2003, pp. 252–259. ACL, Edmonton, Canada (2003)
30. Toutanova, K., Manning, C.D.: 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), Vol. 63–70. ACL, Hong Kong (2000)
31. Klein, D., Manning, C.D.: Accurate Unlexicalized Parsing. In: 41st Meeting of the Association for Computational Linguistics, pp. 423–430. ACL, Sapporo, Japan (2003)
32. Green, S., de Marneffe, M.-C., Bauer, J., Manning, C.D.: 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. ACL, Edinburgh, UK (2010)
33. Porter, M., Boulton, R.: Snowball, <http://snowball.tartarus.org/> (2002)
34. Jadelot, C., Mangeot, M., Petitjean, E., Salmon-Alt, S.: Morphalou 2. In: CNRTL (ed.) (2006)
35. Finkel, J.R., Grenager, T., Manning, C.D.: 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. ACL, Ann Arbor, MI (2005)
36. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4) (2013)
37. Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., Manning, C.D.: A Multi-Pass Sieve for Coreference Resolution. In: Conference on Empirical Methods in Natural Language Processing (EMNLP '10), pp. 492–501. ACL, Cambridge, MA (2010)
38. Miller, G.A.: WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), 39–41 (1995)
39. Sagot, B., Darja, F.: Building a free French wordnet from multilingual resources. In: Ontolex 2008, Marrakech, Maroc (2008)
40. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics, ACL '94, pp. 133–138. ACL, New Mexico, USA (1994)
41. Leacock, C., 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, MA (1998)
42. Denhière, G., Lemaire, B., Bellissens, C., Jhean-Larose, S.: A semantic space for modeling children's semantic memory. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) *Handbook of Latent Semantic Analysis*, pp. 143–165. Erlbaum, Mahwah (2007)
43. Dascalu, M., Trausan-Matu, S., Dessus, P.: Utterances assessment in chat conversations. *Research in Computing Science*, 46, 323–334 (2010)

44. Lemaire, B.: Limites de la lemmatisation pour l'extraction de significations. In: 9es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2009), Lyon, France (2009)
45. Wiemer-Hastings, P., Zipitria, I.: Rules for syntax, vectors for semantics. In: 22nd Annual Conference of the Cognitive Science Society. Erlbaum, Mahwah, NJ (2000)
46. Low, Y., Bickson, D., Gonzalez, J., Guestrin, C., Kyrola, A., Hellerstein, J.M.: Distributed GraphLab: a framework for machine learning and data mining in the cloud. *Proceedings of the VLDB Endowment*, 5(8), 716–727 (2012)
47. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit, <http://mallet.cs.umass.edu/> (2002)
48. Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., Hellerstein, J.M.: GraphLab: A New Parallel Framework for Machine Learning. In: *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 340–349, Catalina Island, California (2010)
49. Dascalu, M., Trausan-Matu, S., Dessus, P.: Cohesion-based Analysis of CSCL Conversations: Holistic and Individual Perspectives. In: 10th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2013), Vol. 1, pp. 145–152. ISLS, Madison, USA (2013)
50. Trausan-Matu, S., Stahl, G., Sarmiento, J.: Supporting polyphonic collaborative learning. *Indiana University Press, E-service Journal*, 6(1), 58–74 (2007)
51. Rebedea, T., Dascalu, M., Trausan-Matu, S., Chiru, C.G.: Automatic Feedback and Support for Students and Tutors Using CSCL Chat Conversations. In: *First International K-Teams Workshop on Semantic and Collaborative Technologies for the Web*, pp. 20–33. Politehnica Press, Bucharest, Romania (2011)
52. Trausan-Matu, S., Rebedea, T.: A Polyphonic Model and System for Inter-animation Analysis in Chat Conversations with Multiple Participants. In: 11th Int. Conf. Computational Linguistics and Intelligent Text Processing (CICLing 2010), Vol. LNCS, pp. 354–363. Springer, Iasi, Romania (2010)
53. Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., Nardy, A.: ReaderBench, an environment for analyzing text complexity and reading strategies. 16th Int. Conf. on Artificial Intelligence in Education (AIED 2013), Memphis (2013)
54. Abrams, E.: Topic Sentences and Signposting. Harvard University, Writing Center (2000)
55. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*, Vol. 1. Cambridge University Press, Cambridge, UK (2008)
56. Galley, M., McKeown, K.: Improving Word Sense Disambiguation in Lexical Chaining. In: 18th International Joint Conference on Artificial Intelligence (IJCAI'03), pp. 1486–1488. Morgan Kaufmann Publishers, Inc., Acapulco, Mexico (2003)
57. Williams, M.: *Wittgenstein, Mind and Meaning: Towards a Social Conception of Mind*. Routledge, New York, NY (2002)
58. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: *International AAAI Conference on Weblogs and Social Media*, pp. 361–362. AAAI Press, San Jose, CA (2009)

59. Brandes, U.: A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology*, 25(2), 163–177 (2001)
60. Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., Nardy, A.: ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. In: 16th Int. Conf. on Artificial Intelligence in Education (AIED 2013). Springer, Memphis, USA (in press)
61. Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp. 404–411. ACL, Barcelona, Spain (2004)
62. Vidal, N.: *Miguel de la faim*. Rageot., Paris (1984)
63. Pfeffer, P.: Les pharmacies des éléphants. In: Pfeffer, P. (ed.) *Vie et mort d'un géant : l'éléphant d'Afrique*, pp. 135. Flammarion, Paris, France (1989)
64. Mandin, S.: *Modèles cognitifs computationnels de l'activité de résumer : expérimentation d'un eiah auprès d'élèves de lycée*. Laboratoire des Sciences de l'Éducation, Vol. Doctoral dissertation. Université Grenoble-2 - Pierre-Mendès-France, Grenoble, France (2009)
65. Donaway, R.L., Drummey, K.W., Mather, L.A.: A comparison of rankings produced by summarization evaluation measures. In: Workshop on Automatic summarization (NAACL-ANLP-AutoSum '00), Vol. 4, pp. 69–78. ACL, Stroudsburg, PA (2000)
66. McNamara, D.S., O'Reilly, T.P., Rowe, M., Boonthum, C., Levinstein, I.B.: iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In: McNamara, D.S. (ed.) *Reading comprehension strategies: Theories, interventions, and technologies*, pp. 397–420. Erlbaum, Mahwah, NJ (2007)
67. Dahl, R.: *Matilda*. Gallimard, Paris, France (2007)
68. Graesser, A.C., Singer, M., Trabasso, T.: Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3), 371–395 (1994)
69. Dascalu, M., Trausan-Matu, S., Dessus, P.: Towards an integrated approach for evaluating textual complexity for learning purposes. In: 11th Int. Conf. in Advances in Web-Based Learning (ICWL 2012), Vol. LNCS 7558, pp. 268–278. Springer, Sinaia, Romania (2012)
70. Cortes, C., Vapnik, V.N.: Support-Vector Networks. *Machine Learning*, 20(3), 273–297 (1995)
71. François, T., Miltakaki, E.: 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. ACL, Montreal, Canada (2012)
72. Petersen, S.E., Ostendorf, M.: A machine learning approach to reading level assessment. *Computer Speech and Language*, 23, 89–106 (2009)
73. van Dijk, T.A., Kintsch, W.: *Strategies of discourse comprehension*. Academic Press, New York, NY (1983)
74. Feng, L., Jansche, M., Huenerfauth, M., Elhadad, N.: A Comparison of Features for Automatic Readability Assessment. In: 23rd Int. Conf. on Computational Linguistics (COLING 2010), Vol. Poster, pp. 276–284. ACL, Beijing, China (2010)
75. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D.: 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. ACL, Portland, OR (2011)

76. Geisser, S.: Predictive inference: an introduction. Chapman and Hall, New York, NY (1993)
77. Winne, P.H., Baker, R.S.J.d.: The potentials of educational data mining for researching metacognition, motivation and Self-Regulated Learning. *Journal of Educational Data Mining*, 5(1), 1–8 (2013)
78. Bouhineau, D., Luengo, V., Mandran, N., Toussaint, B.M., Ortega, M., Wajeman, C.: Open platform to model and capture experimental data in Technology Enhanced Learning systems. *Workshop Data Analysis and Interpretation for Learning Environments*. Alpine Rendez-Vous, Villard-de-Lans (2013)