HAL

archives-ouvertes.fr

Are Automatically Identified Reading Strategies Reliable Predictors of Comprehension?

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

▶ To cite this version:

HAL Id: hal-01390231

https://hal.archives-ouvertes.fr/hal-01390231

Submitted on 1 Nov 2016

Are Automatically Identified Reading Strategies Reliable Predictors of Comprehension?

Mihai Dascalu[1], Philippe Dessus[2], Maryse Bianco[2], Stefan Trausan-Matu[1]

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

Abstract. In order to build coherent textual representations, readers use cognitive procedures and processes referred to as reading strategies; these specific procedures can be elicited through self-explanations in order to improve understanding. In addition, when faced with comprehension difficulties, learners can invoke regulation processes, also part of reading strategies, for facilitating the understanding of a text. Starting from these observations, several automated techniques have been developed in order to support learners in terms of efficiency and focus on the actual comprehension of the learning material. Our aim is to go one step further and determine how automatically identified reading strategies employed by pupils with age between 8 and 11 years can be related to their overall level of understanding. Multiple classifiers based on Support Vector Machines are built using the strategies' identification heuristics in order to create an integrated model capable of predicting the learner's comprehension level.

Keywords: Self-Explanations, Reading Strategies, Comprehension Prediction, Identification Heuristics, Support Vector Machines.

1   Introduction

In order to build textual coherence and to achieve a consistent representation of the discourse, readers need to transcend beyond what is explicitly expressed by employing cognitive procedures and processes, referred to as reading strategies. Those procedures are elicited through self-explanations [1]. Research on reading comprehension has shown that expert readers use specific strategies to on-line monitor their reading, thus being able to know at every moment their level of understanding. Moreover, when faced with a difficulty, learners can call upon regulation procedures, also part of reading strategies [2]. In this context, psychological and pedagogical research has revealed that people tend to understand better a text if they try to explain themselves what they have read [3]. Starting from these observations, techniques such as *SERT* (Self-Explanation Reading Training) [4], were developed to support students better understand texts.

Reading strategies have been extensively studied with adolescent and adult readers using the think-aloud procedure that engages the readers to self-explain what they understood so far at specific breakpoints while reading. Our study is focused on comprehension assessment for an audience more rarely studied, primary pupils, whose guidance plays a central role. As previous research suggests, self-regulation can be enhanced through the use of metacognitive reading strategies [5]. Pupils tend to better understand a given text by employing these specific mechanisms [6]. Also, this paper represents a continuation of previous research [7, 8], with a refined set of heuristics for best matching human annotations, accompanied by a prediction mechanism based on Support Vector Machines [9] in order to estimate pupil's comprehension level of a given text.

The following section presents an overview of the evaluation of reading strategies, their categorization, and other similar automated systems that have been developed to identify the employed reading strategies. The third section is centered on the description of the used heuristics, while the fourth section introduces the classifier that combines the previously identified reading strategies and predicts the learner's comprehension level. Afterwards, the fifth section encompasses the performed validations for testing the system's accuracy, while the last section is focused on conclusions and future improvements.

2 Overview of the Assessment of Reading Strategies

Expert readers frequently make use of four types of reading strategies in order to achieve a deep understanding from the texts they read [4]. *Paraphrasing* enables readers to express what they understood from the explicit content of the text and can be considered the first and essential step in order to achieve a coherent representation. *Text-based inferences*, consisting predominantly of causal and bridging strategies, build explicit relationships between two or more textual segments of the initial text. On the other hand, *knowledge-based inferences* create relationships between the information from the text and the reader's personal knowledge and are essential to create the situation model [10]. *Control strategies* refer to the actual monitoring process, when readers explicitly express what they have or have not understood.

Nevertheless, if we want students to be assisted while reading, one human expert (e.g., a teacher) can take care only after a small number of them, which makes it impossible for such training techniques to be used on a large scale. For example, this is one of the major problems of MOOCs (Massively Online Open Courses) in which, due to the previous constraints, assistance is frequently provided by peer students, increasing nevertheless the risk of making mistakes [11, 12]. Moreover, assessing the content of a verbalization is a demanding and a subjectivity-laden activity, which can be assisted by computer-based techniques. These are the main motives behind the idea of using a computer program instead of, or as support for, a human tutor.

Initial experiments were conducted by McNamara and her colleagues [13] and *iSTART* [14] can be considered the first implemented system that addresses self-explanations [15]. It has various modules that explain the *SERT* method to the

students, one that shows them how to use those techniques using a virtual student, and another training module that asks students to read texts and give verbalizations, evaluates them and provides an appropriate feedback. *iSTART* divides verbalizations into four main categories: irrelevant, paraphrases, verbalizations that use knowledge previously found in the text and verbalizations which use external knowledge from the students' experience. It is easier to automatically identify paraphrases and irrelevant explanations, but it is more difficult to identify and evaluate verbalizations that contain information coming from students' experience [16].

We conducted an experiment [6] for analyzing the control and the regulation of comprehension through reading strategies. Pupils (3rd–5th grade, 8–11 years old) were given the task to read aloud two French stories and were asked at predefined moments to self-explain their impressions and thoughts about the reading materials. The self-explanations were coded according to McNamara's [4] scheme. The results of this study support the view that pupil's self-explanations are an adequate way to access to their reading strategies. The sole exception consists of prediction strategies, which were scarcely used in comparison to McNamara's participants, perhaps due to the age of the pupils. Initial and partial automated results based on the previous study were presented in [7], and we present in this paper data from a larger sample, using fine-tuned heuristics and an automatic classifier for predicting comprehension.

3 Reading Strategies Identification Heuristics

In terms of reading strategies, our aim was to create automated extraction methods designed to support tutors at identifying various strategies employed by pupils that are best aligned with the annotation categories: 1/ *paraphrasing*, 2/ *text-based inferences* consisting of *causality* and *bridging*, 3/ *knowledge-based inferences* or elaboration and 4/ monitoring or *control* [6]. A clear demarcation between causal inferences and bridging had to be established within our automated system due to underlying approaches and computational complexity, although causal inferences can be considered a particular case of bridging, as well as a reference resolution. In addition, we have tested various methods of identifying reading strategies and we will focus solely on presenting the refined heuristics that provided in the end the best overall human–machine correlations.

In ascending order of complexity, the simplest strategies to identify are *causality*, with markers like "*parce que*" (because), "*pour*" (for), "*donc*" (thus), "*alors*" (then), "*à cause de*" (because of) and *control*, with markers like "*je me souviens*" (I remember), "*je crois*" (I believe that), "*j'ai rien compris*" (I haven't understood anything) for which cue phrases based on pattern matching techniques have been used. As particular refinement for causality, all occurrences of the keywords at the beginning of a verbalization have been discarded because the strategy needs to create an inferential link between two adjacent textual segments, out of which the first is lacking since it is the beginning of a verbalization. In this particular case, the use of causality patterns indicates a lacunar pupil formulation frequently observed at their age. In terms of control, besides the verification of specific cue phrases, we added a

check to verify whether the pattern exists in the sentences within the original text, in which case we would be dealing with a paraphrase rather than a control statement.

As a second stage of complexity, *paraphrases*, that in the manual annotation were considered mere repetitions of the same semantic propositions by human raters, were automatically identified through lexical similarities. More specifically, words from the verbalization were considered as paraphrased words if they had identical lemmas or stems, or were synonyms extracted from lexicalized ontologies – *WordNet* [16] or *WOLF* [17] – with words from the initial text. Adjacent words from pupil's self-explanations, identified as paraphrased concepts were grouped into paraphrase segments in order to highlight contiguous zones highly referential to the initial text. In addition, if more than a predefined percentage of relevant words from a sentence from the initial text are paraphrased within the verbalization, that specific sentence is tagged as a paraphrasing segment. The previous percentage was empirically set after performing multiple iterations with incremental values, whereas relevant words are obtained after stop words elimination and after selecting solely dictionary words.

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 highest similarity to another word from the initial text (expressed in terms of semantic distances in ontologies, Latent Semantic Analysis and Latent Dirichlet Allocation) [7] and the relevance of both words to the textual fragments in-between consecutive self-explanations expressed as semantic cohesion. 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 semantic similarities 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.

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 [7]. Semantic similarity was measured in-between the current verbalization and the two previous textual blocks from the initial text. In order to relate to the overall cohesion between the verbalizations and what was initially stated within the reading material, the imposed similarity threshold for tagging a sentence as being a bridged element uses a cohesion value that exceeds the mean plus standard deviation of all previous similarity measures performed on all self-explanations of a given pupil. Similarly to paraphrases and for best adapting to the manual annotation process, adjacent sentences from the initial text tagged as being bridged within the verbalization are grouped into a bridging segment. Moreover, if a sentence is considered to be a paraphrasing segment due to a high density of paraphrased words, that sentence is not taken into consideration while defining the final bridging segments. To better highlight the identification mechanisms, Fig. 1 depicts with bold bridged sentences from the initial text with verbalization 2 that exceed the identified threshold and that are not marked as paraphrases. In the end, four bridged segments are automatically determined: A3, B1 together with B2 due to adjacency within the same paragraph, C1 and C3 from the later textual block.
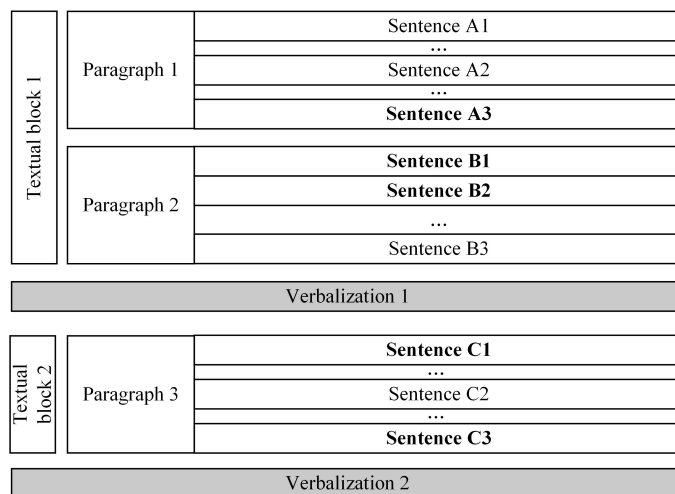
		Sentence A1
	Paragraph 1	...
		Sentence A2
Textual block 1		...
		Sentence A3
		Sentence B1
	Paragraph 2	**Sentence B2**
		...
		Sentence B3
Verbalization 1		
		Sentence C1
Textual block 2	Paragraph 3	...
		Sentence C2
		...
		Sentence C3
Verbalization 2		

Fig. 1. Bridging identification.

4 Combining Automatically Identified Reading Strategies for Predicting Comprehension

All the previous reading strategies and their corresponding identification heuristics can be viewed as attributes that describe the learner's comprehension level. In order to *predict* the comprehension level of each learner based on the used reading strategies, post-tests were administered to each pupil and comprehension scores were manually determined using these tests. Therefore, we found it appropriate to use a classifier that accepts as inputs the number of used reading strategies and predicts a comprehension class depicting the reader's understanding level expressed as a comprehension level class estimate.

Similar to the textual complexity problem for which Support Vector Machines (SVMs) [9] have been proven to be the most relevant [18], we trained multiple SVMs for determining the appropriate comprehension class. A one-versus-all approach implementing the winner-takes-all strategy is used to deal with the problem of multiple SVM returning 1 for a specific text (the classifier with the highest output function assigns the class). As specific optimizations, an RBF kernel with degree 3 was selected and a Grid Search method [19] was enforced to increase the effectiveness of the SVM through the parameter selection process for the Gaussian kernel. Exponentially growing sequences for C and γ were used, and each combination of parameter choices was checked using the testing corpora; in the end, the parameters that generated the best precision were selected.

5 Validations of the Identification Heuristics and of the Comprehension Prediction Model

We ran an experiment with 82 pupils with age between 8 and 11 years, uniformly distributed in terms of their age, who had each to read aloud two French stories of about 450 words (*The Cloud Swallower* and *Matilda*). During their lecture, pupils had to stop in-between at five, respectively 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 in the end categorized according to the imposed annotation scheme. Nevertheless, when looking at manual assessments, discrepancies between evaluators were identified due to different understandings and perceptions of pupil's intentions, expressed within their self-explanations; all disagreements were solved individually by mediation for each self-explanation. In addition, predefined rules and patterns were used to perform automatic cleaning in order to process the phonetic-like transcribed verbalizations.

Fig. 2. Visualization of automatically identified reading strategies.

Fig. 2 depicts the main interface of our developed system in which the grey sections represent the pupil's self-explanations, whereas the white blocks represent paragraphs from the read story. All strategies are highlighted within the self-explanation with a specific color encoding: control, causality, paraphrasing, inferred concept [*] and bridging with a clear demarcation of the textual segments from the reading material comprising of inter-linked cohesive sentences. In addition, Fig. 2 also depicts in the last column the cohesion measures normalized in [0; 1] with previous paragraphs from the story.

Three variables were required for fine-tuning the higher-level reading strategies: bridging requires a *minimum semantic cohesion* ($Min_{coh\_bridging}$) and a *maximum percentage* of words for not considering a sentence as paraphrased ($Max_{paraphrase}$), while knowledge inference uses only a *minimum similarity threshold* (Min_{sim_KI}). Our system automatically determines the most suitable values for maximizing the overall Pearson correlations and *F1*-scores as measures of outputs' correctness with regards to the manual annotations (see Table 1). As expected, paraphrases, control and causality occurrences were much easier to identify than information coming from pupils' experience [20]. Moreover, our experiments demonstrate that although the variables for the two texts have similar optimal values, there are rather high fluctuations in the accuracy of the reading strategies' identification, therefore highlighting the specificities of each text and the intrinsic subjectivity of the analysis.

Table 1. Accuracy of the automatically identified reading strategies.

Statistic measure	Paraphrasing	Text-based Inference (causality and bridging)	Knowledge-based Inference	Control
Text 1: *The Cloud Swallower*				
$Min_{coh\_bridging}$ = .40; $Max_{paraphrase}$ = 60%; Min_{sim_KI} = .33				
Pearson correlation	.64	.55	.41	.84
Precision	.64	.79	.50	.76
Recall	.99	.83	.94	.63
F1 score	.78	.81	.65	.68
Text 2: *Matilda*				
$Min_{coh\_bridging}$ = .45; $Max_{paraphrase}$ = 65%; Min_{sim_KI} = .33				
Pearson correlation	.56	.69	.48	.90
Precision	.73	.71	.34	.86
Recall	.99	.94	.97	.70
F1 score	.84	.81	.50	.77
All verbalizations together, from both texts				
$Min_{coh\_bridging}$ = .4; $Max_{paraphrase}$ = 65%; Min_{sim_KI} = .33				
Pearson correlation	.64	.60	.35	.89
Precision	.69	.74	.47	.83
Recall	.99	.90	.87	.68
F1 score	.81	.81	.61	.74

After fine-tuning the identification heuristics, we opted to create three comprehension classes for predicting the learner's comprehension level with a

distribution of 30%, 40% and 30% of all pupil scores sorted in ascending order and to apply 3-fold cross-validations for the SVM training process. The resulting average agreement between automatic predictions and the class assigned from the post-test scores was approximately .78 in most runs (see Table 2). Due to a rather limited corpus, the prediction accuracy oscillates between different training sessions, with a minimum of .66. We also noticed a rather small differentiation between the first and the second class, as well as conflicting instances of pupils with a high number of used reading strategies, but pertaining to opposite comprehension classes. The previous contradictions in terms of the number of used reading strategies in opposition to pupils' comprehension levels, corroborated with rather small differentiations between adjacent classes, led to a rather low prediction accuracy of the second class.

Table 2. Comprehension prediction based solely on the four automatically identified reading strategies.

Verbalizations pertaining to	Agreement – Class 1 –	Agreement – Class 2 –	Agreement – Class 3 –	Average agreement
Text 1: *The Cloud Swallower*	1	.33	.67	.67
Text 2: *Matilda*	.67	.33	1	.67
Both texts	1	.33	1	.78

Nevertheless, results are encouraging based on the limited number of training instances, the reduced number of classification attributes and the fact that a lot of noise existed within the transcriptions. From this point, it becomes clear that external factors should be enforced in order to increase the accuracy of the prediction and to create a more comprehensive view, as the diversity and the richness of the strategies a reader carries out depend on many factors, either personal (proficiency, level of knowledge, motivation), or external (textual complexity).

In order to prove the feasibility of the previous statements, we added a simple factor already computed during the identification process: the average value of cohesion between each verbalization and the corresponding paragraphs from the initial text. This measure emphasizes the link between what was initially stated and the learner's understanding or personal perspective. As expected, the results from Table 3 highlight an increase in the overall prediction accuracy.

Table 3. Comprehension prediction based on the four heuristics plus the average cohesion value added as an attribute for classification.

Verbalizations pertaining to	Agreement – class 1 –	Agreement – class 2 –	Agreement – class 3 –	Average agreement
Text 1: *The Cloud Swallower*	1	.33	1	.78
Text 2: *Matilda*	1	.67	.67	.78
Both texts	1	.67	1	.89

In the end, notable improvements in terms of the initial experiments presented in [7] can be observed: 1/ the use of 8 times more participants, each self-explaining two texts instead of only one; 2/ an important increase in the identification accuracy for

paraphrases, knowledge and text-based inferences; 3/ although bridging taken individually has still a low correlation which indicates that the human annotated bridging strategy is not aligned with the identification heuristics, the use of the new class of text-based inferences demonstrates that the integrated perspective of bridging and causality taken together is more cognitively relevant and representative with regards to the manual annotations; 4/ the use of Support Vector Machines for predicting the learner's comprehension level.

6　Conclusion and Future Research Directions

Our aim consists of supporting tutors and our approach emphasizes the benefits of a regularized and deterministic process of identification as a viable alternative to the subjectivity-laden task of manual annotation. Moreover, the performed validations confirm that reading strategies are related to the pupil's comprehension level, but also highlight the need to add more factors, potentially inspired from textual complexity measures [21, 22] or essay scoring techniques [23] in order to increase the accuracy of the predictions.

As the comprehension scores are not global, but related to the read texts subject to expressing one's meta-cognitions, we can state that reading strategies can be used to predict comprehension based on the overall experimental settings. Our next aim consists of deploying and using our system in classroom settings to analyze student's reading strategies and to infer possible comprehension problems in near realtime.

Acknowledgements

References

1. Millis, K., Magliano, J.P.: Assessing comprehension processes during reading. In: Sabatini, J.P., Albro, E.R., O'Reilly, T. (eds.) Assessing reading in the 21st century, pp. 35–53. Rowman & Littlefield Publishing, Lanham, MD (2012)
2. 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, NJ (2009)
3. McNamara, D.S., Scott, J.L.: Training reading strategies. 21th Annual Meeting of the Cognitive Science Society (CogSci '99), pp. 387–392. Erlbaum, Hillsdale (1999)
4. McNamara, D.S.: SERT: Self-Explanation Reading Training. Discourse Processes, 38, 1–30 (2004)

5. Nash-Ditzel, S.: Metacognitive Reading Strategies Can Improve Self-Regulation. Journal of College Reading and Learning, 40(2), 45–63 (2010)

6. 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, pp. 16. Peter Lang, Bern-Paris (in press)

7. Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., Nardy, A.: ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. In: AIED 2013, Vol. LNCS 7926, pp. 379–388. Springer, Memphis, USA (2013)

8. Dascalu, M.: Analyzing Discourse and Text Complexity for Learning and Collaborating, Studies in Computational Intelligence, Vol. 534. Springer, Switzerland (2014)

9. Cortes, C., Vapnik, V.N.: Support-Vector Networks. Machine Learning, 20(3), 273–297 (1995)

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

11. Piech, C., Huang, J., Chen, Z., Do, C., Koller, D.: Tuned models of peer assessment in MOOCs. In: Int. Conf. Educational Data Mining (EDM 2013). International Educational Data Mining Society, Memphis, TN (2013)

12. Goldin, I.M.: Acounting for peer reviewer bias with Bayesian models. In: ITS 2012, Vol. LNCS 7315. Springer, Chania, Grece (2012)

13. O'Reilly, T.P., Sinclair, G.P., McNamara, D.S.: iSTART: A Web-based Reading Strategy Intervention that Improves Students' Science Comprehension. In: CELDA2004, pp. 8. IADIS Press, Lisbon, Portugal (2004)

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

15. Jackson, G.T., Guess, R.H., McNamara, D.S.: Assessing cognitively complex strategy use in an untrained domain. In: 31st Annual Meeting of the Cognitive Science Society (CogSci '09), pp. 2164–2169. Cognitive Science Society, Amsterdam, The Netherlands (2009)

16. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM, 38(11), 39–41 (1995)

17. Sagot, B.: WordNet Libre du Francais (WOLF), http://alpage.inria.fr/~sagot/wolf.html (2008)

18. François, T., Miltsakaki, E.: Do NLP and machine learning improve traditional readability formulas? In: PITR2012, pp. 49–57. ACL, Montreal, Canada (2012)

19. Bergstra, J., Bengio, Y.: Random Search for Hyper-Parameter Optimization. The Journal of Machine Learning Research, 13, 281–305 (2012)

20. Graesser, A.C., Singer, M., Trabasso, T.: Constructing inferences during narrative text comprehension. Psychological Review, 101(3), 371–395 (1994)

21. Graesser, A.C., McNamara, D.S., Kulikowich, J.: Coh-Metrix: Providing multilevel analyses of text characteristics. Educational Researcher, 40(5), 223–234 (2011)

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

23. Todd, R.W., S.Khongput, Darasawang, P.: Coherence, cohesion and comments on students' academic essays. Assessing Writing, 12(10–25) (2007)