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

As an overview, in terms of individual learning, *ReaderBench* encompasses the functionalities of both *CohMetrix* (McNamara et al. 2010) (see *2.2.2* Textual Complexity Computational Approaches) and *iStart* (McNamara et al. 2007a; Graesser et al. 2005) (see *2.3* Reading Strategies), as it provides teachers and learners information on their reading/writing activities: initial textual complexity assessment, assignment of texts to learners, capture of metacognitions reflected in one's textual verbalizations, and reading strategies assessment (a detailed comparison is presented at the end of this chapter). Moreover, *ReaderBench* encompasses textual complexity measures similar to *Dmesure* (François and Miltsakaki 2012; François 2012), but with emphasis on more in-depth, semantic factors. The main differentiators between *ReaderBench* and the previous systems consist of the following (see *8.3* Comparison of *ReaderBench* to *iSTART, Dmesure and Coh-Metrix* for more details):

- Emphasis on comprehension extracted from the automatic analysis of metacognitions (Dascalu et al. 2013a), based on two preliminary studies (Oprescu et al. in press; Dessus et al. 2012).
- A different educational purpose, as *ReaderBench* validation was performed on primary school pupils, whereas *iStart* mainly targets high school and university students (Nardy et al. in press).
- Different factors, measurements and the use of SVMs (Cortes and Vapnik 1995; François and Miltsakaki 2012) for increasing the validity of textual complexity assessment (Dascalu et al. 2012).

8.1 Identification of Reading Strategies

The use of reading strategies is widely recognized as a crucial determinant of reading comprehension (see *2.3* Reading Strategies). Second degree and high school pupils who are good comprehenders are mostly strategic readers (Graesser

M. Dascalu, *Analyzing Discourse and Text Complexity for Learning and Collaborating*, Studies in Computational Intelligence 534,
DOI: 10.1007/978-3-319-03419-5_8, © Springer International Publishing Switzerland 2014

2007). These strategies can be elicited through self-explanations (Chi et al. 1994) and have been categorized by McNamara (2004) as follows: comprehension monitoring, paraphrasing, elaboration, prediction, and bridging. One important skill that these strategies exploit is to be able to establish semantic and causal relationships between the read sentences (Wolfe et al. 2005).

Based on these findings, McNamara et al. (2007a) developed *iSTART*, a cognitive tutor that automatically categorizes self-explanations, partly using Latent Semantic Analysis (Landauer and Dumais 1997). Any thorough analysis of self-explanations reports it is a very demanding and subjectivity-oriented activity, and the use of systems like *iSTART* to detect pupils' reading strategies is more than challenging. Since a cognitive tutor guides the reader through predefined steps alternating between reading and verbalizations, a similar computer-based scenarization is made possible through the wide range of reading strategies and the feedback possibilities (Vitale and Romance 2007). Nevertheless, as our focus was to automatically assess verbalizations and to identify reading strategies, multiple alternatives were explored: two initial studies addressed in extent the identification of paraphrases (Dessus et al. 2012; Oprescu et al. in press), while an integrated view targeting the automatic identification of all proposed reading strategies (both low-level – causality, control, paraphrasing – and high-level, cognitive strategies – knowledge inference and bridging) was first introduced in *ReaderBench* (Dascalu et al. 2013a).

The *data gathering* and *evaluation method* applied a priori was the same for all experiments, but the corpus of evaluated verbalizations consisted of different sub-sets of the entire collection. In the end, during the ANR DEVCOMP project, 84 pupils from 3rd to 5th grades, from the same school and from a middle socio-economic background participated in our experiments. The pupils read a narrative text consisting of 453 words, the story "*Matilda*" by Dahl (2007), and explained what they understood up to that point at 6 predefined breakpoints (see Appendix D – Input Examples, Sample Document – Matilda by for complete text). The text was chosen to be within the reading level of participants, so that differences in verbalizations would indicate differences in reading strategies instead of comprehension difficulties. In order to perform a fine-grained analysis, the initial text was split in 45 segments (of about 1 sentence each). A causal analysis was performed so that both local (when the causal antecedent is close to the reference sentence) and distal antecedents (when the causal antecedent is somewhat farther, out of the reader's working memory) of sentences were determined in accordance to Millis et al. (2006). Finally, a propositional analysis of the text was proposed that allowed us to extract macro-propositions and to support the coding of what was remembered by the participants.

Participants individually read the text out loud and stopped at predetermined breaks to self-explain the text segment just read, the whole activity being recorded. The task was explained to pupils as follows: "During your reading you will stop at each icon to tell out loud what you have understood, just at this time". Their verbalizations were then transcribed and each self-explanation was semantically compared using different natural language processing techniques. Pupils' verbalizations were analyzed proposition by proposition and were categorized by

experts according to a coding scheme adapted from McNamara (2004). Disagreements between experts in terms of identified reading strategies were discussed and resolved by consensus (Nardy et al. in press). As technical specificity, the first two studies were conducted using LSA vector spaces trained on the "*TextEnfants*" corpus (Denhière et al. 2007) (approx. 4.2 M words) with no specific NLP or Information Retrieval optimizations (only stop words elimination), while *ReaderBench* also integrated "*Le Monde*" corpus (French newspaper, approx. 24 M words) with all optimizations mentioned in *7.2* Cohesion-based Discourse Analysis.

8.1.1 The Initial Study of Analyzing Paraphrases

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

Our research questions were: 1/ to compare human expert categorization of paraphrases to the semantic similarity between text sentences and self-explanations, obtained by means of LSA; 2/ to highlight an expected "recency effect", stating that the information children self-explain most often pertains to very close sentences to the verbalization break; 3/ to investigate the way pupils account for causal relations (either local or distal) in retelling causally related text sentences.

Firstly, we computed accuracy measures in order to compare human vs. LSA values of sentence relatedness and to check the validity of the computer-based measures. Pearson correlations between the number of paraphrases per verbalization (*Vn*) detected by the two raters and LSA similarities between each verbalization and the previous sentences were as follows: *V1*: $r = .48$; *V2*: $r = .58$; *V3*: $r = .74$; *V4*: $r = .29$; *V5*: $r = .57$; *V6*: $r = .61$, which shows that human judgments of paraphrases expressed by children on each paragraph are moderately to strongly related to LSA measures of similarities.

Secondly, we investigated the extent to which each self-explanation was related to the last read sentence (focal) (see Figure 41). We observed that the recency effect varies across verbalization plots, indicating that this effect is dependent of the content conveyed by the last sentences. Moreover, the focal sentence, in general, does not have a higher similarity with the related verbalization than the average of other previous sentences, except for *V4*: $t(43) = 7.5$, $p < .0005$. Two-way ANOVAs showed a significant difference between grades for *V6*, $F(1, 42) = 7.01$; $p < .05$ and a tendency for *V2*, $F(1, 42) = 3.22$, $p < .09$. Although grade 3 pupils tended to recall the last sentence at these points more frequently, the semantic content of the last sentence seems to be the main determinant of focal recall.

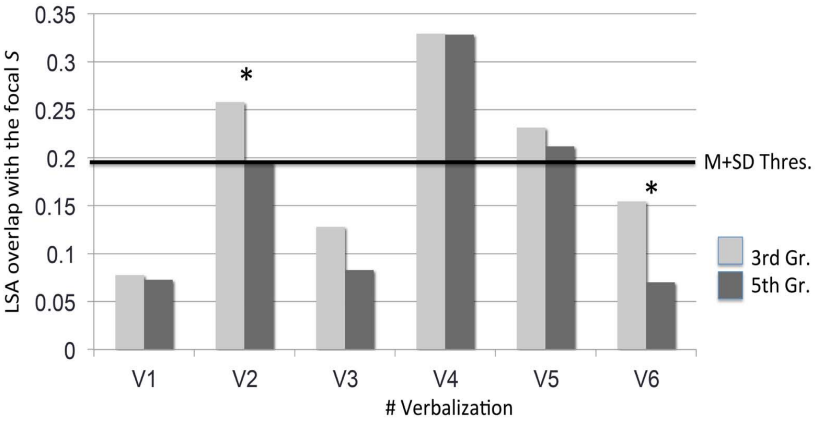


Fig. 41 *Matilda* – Mean LSA-based values for similarity of focal sentences by grade

Thirdly, we initially expected that 1/ the semantic content of local and distal sentences, as determined by the causal analysis, is more often verbalized than the rest of the previous text and the focal sentence and 2/ the local-centered causal sentences are better recalled than the distal-centered ones (see Figure 42). Results first showed that local and distal causal sentences are, in all cases but two (local vs. V_1 and V_5), significantly more verbalized than the rest of the text. Moreover, the content of local causal sentences was significantly better recalled than focal sentences in V_1 and V_3 (resp. $t(43) = 3.11$, $p < .005$; $t(43) = 9.45$, $p < .0005$). Unexpectedly, the content of distal causal sentences was better recalled than local causal sentences for V_1: $t(43)=6.09$, $p < .0005$; V_2: $t(43)=8.49$, $p < .0005$. Two-way

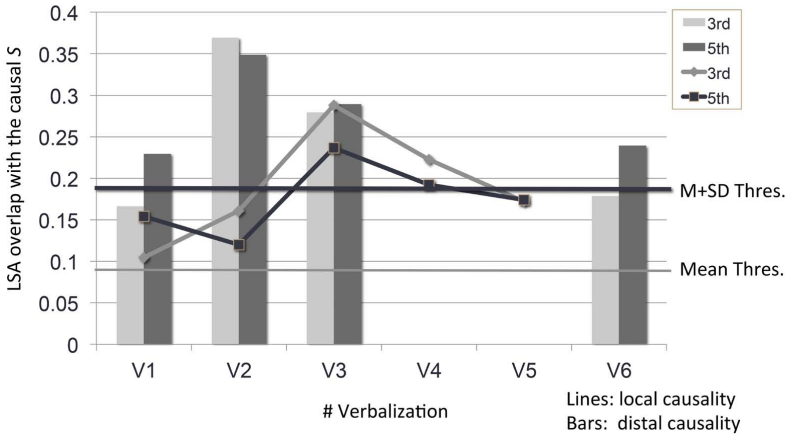


Fig. 42 *Matilda* – Mean LSA-based values for similarity of causal sentences, by grade. Lines: local causality; bars: distal causality

ANOVAs showed significant differences between grades for V_1 (distal), $F(1, 42) = 4.43$, $p < .05$; and a tendency for V_6 (distal), $F(1, 42) = 3.90$, $p < .06$ and for V_3 (local), $F(1, 42) = 2.91$; $p < .1$. Overall, participants' strategies focused on causality, rather than recency.

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

8.1.2 The Second Study of Analyzing Paraphrases

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

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

where RR_1 is the rating returned by the function, nn_1, nn_1, nn_{1r} and nn_{1r} are the number of nouns, verbs, adjectives and respectively adverbs in the verbalization that can be found in the list of relevant nouns of the paragraphs, NN_1, NN_1, NN_{1r} and NN_{1r} are the length of these lists and, and WW_1,WW_1, WW_{1r} and WW_{1r} are their weights in the average. All these predefined weights were determined experimentally, after running multiple iterations with incremental values.

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

At this point we had introduced two metrics, both indicating the degree of resemblance of two paragraphs, but we had to decide whether the results of these two metrics are coherent or not, so we tried to evaluate the correspondence between the two metrics (see Figure 43). Based on these observations, we decided that the best way to combine these two metrics was to multiply them. The combined metrics is also represented in the same chart.

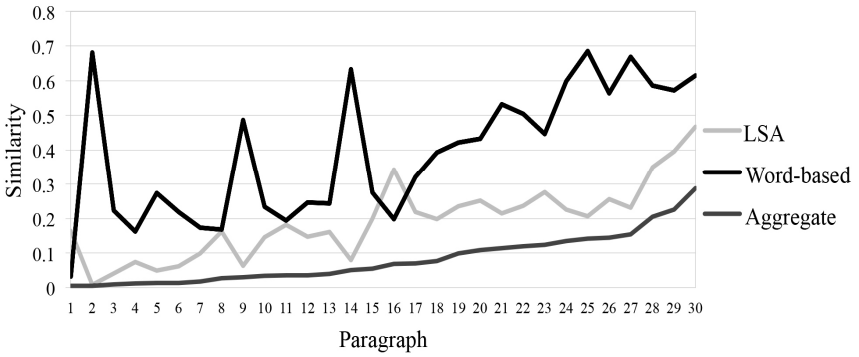


Fig. 43 Comparison between the LSA similarity and word-based heuristics

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

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

Figure 44 shows the values returned by the word-based metrics for ten paraphrases, which represent about one third of the total number of paraphrases of our test corpus, when compared to the previous, the current and the next segment of text that consists of a merge of all paragraphs in-between two adjacent self-explanations. It is obvious that there is higher resemblance between the current textual segment and the verbalization (so the one just in front of the metacognition break, recalled by the pupil), while the similarity between the verbalization and other surrounding textual segments is close to zero. There are some exceptions, mainly highlighting different types of verbalizations, but no straightforward conclusion could be drawn and further experiments needed to be conducted.

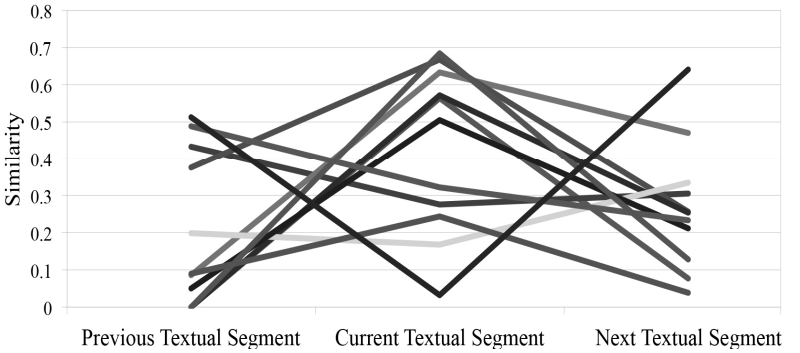


Fig. 44 Comparison of verbalizations containing paraphrases, using the word-based heuristic

Figure 45 depicts a similar analysis using the LSA similarity function. We notice that the graphic has the same characteristics as Figure 44, a similar coefficient of variation ($c_v = 0.7$), but follows more strictly the pattern of positive slope followed by a negative one, which led us to conclude that the LSA method is more accurate than word-based heuristic, although the average similarity values were quite low. Therefore, in this second study we used LSA and a word-based heuristic to compare the verbalizations with nearby paragraphs and this approach provided encouraging results, as we were able to identify paraphrases with good precision. As conclusions, we decided to focus on extracting reading strategies only by comparing the verbalizations to the previous blocks of texts, in-between the previous and the current verbalization. Moreover, the combination of semantic distances from ontologies and LSA seemed a good practice that lead to the aggregated cohesion function integrated in *ReaderBench*.

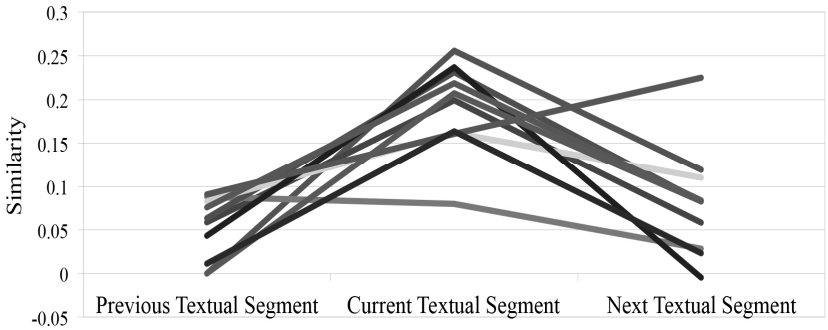


Fig. 45 Comparison of verbalizations containing paraphrases, using the LSA-based heuristic

8.1.3 Reading Strategies Identification Heuristics

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

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

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

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

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

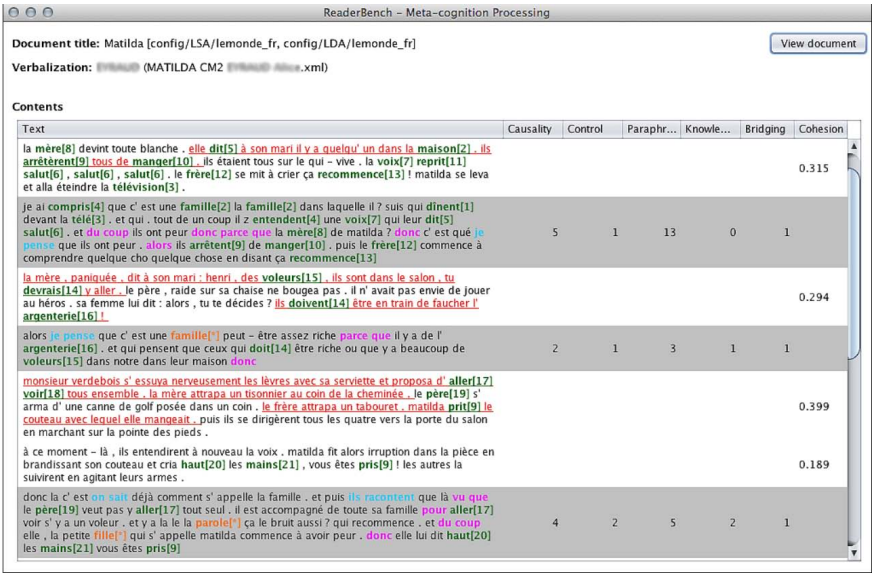


Fig. 46 *ReaderBench* (2) Visualization of automatically identified reading strategies.The grey sections represent the pupil's self-explanations, whereas the white blocks represent paragraphsfrom "Matilda" by Dahl (2007). Causality, control and inferred concepts (that through their definition are not present within the original text) are highlighted only in the verbalization, whereas paraphrasesare coded in both the self-explanation and the initial text for a clear traceability of lexical proximity oridentity. Bridging, if present, is highlighted only in the original text for pinpointing out the textualfragments linked together through cohesion in the pupil's meta-cognition.

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 F1score 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) (Dascalu et al. in press). As expected, paraphrases, control and causality occurrences were much easier to identify than information coming from pupils' experience (Graesser et al. 1994).

Figure 46 depicts the cohesion measures with previous paragraphs from the story in the last column and the identified reading strategies for each verbalization marked in the grey areas, coded as follows: control, causality, paraphrasing [index referred word from the initial text], inferred concept [*] and bridging over the inter-linked cohesive sentences from the reading material. The initial text of the verbalization, including the corresponding manual coding scheme, can be found in Appendix D – Input Examples, Sample Verbalization.

Moreover we have identified multiple particular cases in which both approaches (human and automatic) covered a partial truth that in the end is subjective to the evaluator. For instance, many causal structures close to each other, but not adjacent, were manually coded as one, whereas the system considers each of them separately. For example, "*fille*" ("daughter") does not appear in the text and is directly linked to the main character, therefore marked as an inferred concept by *ReaderBench*, while the evaluator considered it as a synonym. Additionally, when looking at manual assessments, discrepancies between evaluators were identified due to different understandings and perceptions of pupil's intentions expressed within their metacognitions. Nevertheless, our aim was to support tutors and the results are encouraging (correlated also with the previous precision measurements and with the fact that a lot of noise existed in the transcriptions), emphasizing the benefits of a regularized and deterministic process of identification.

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

8.2 Textual Complexity Analysis Model

Assessing textual complexity can be considered a difficult task due to different reader perceptions primarily caused by prior knowledge and experience, cognitive capability, motivation, interests or language familiarity (for non-native speakers) (see *2.1.3* Cohesion and Coherence versus Textual Complexity and *2.2* Textual Complexity). Nevertheless, from the tutor perspective, the task of identifying

accessible materials plays a crucial role in the learning process since inappropriate texts, either too simple or too difficult, can cause learners to quickly lose interest.

In this context, we propose a multi-dimensional analysis of textual complexity, covering a multitude of factors integrating classic readability formulas, surface metrics derived from automatic essay grading techniques, morphology and syntax factors (Dascalu et al. 2012), as well as new dimensions focused on semantics (Dascalu et al. 2013a). In the end, subsets of specific factors are aggregated through the use of Support Vector Machines (Cortes and Vapnik 1995), which has proven to be the most efficient method (François and Miltsakaki 2012; Petersen and Ostendorf 2009). In order to provide an overview, the textual complexity dimension, with their corresponding performance scores, are presented in Table 27, whereas the following subsections describe each dimension with its complexity factors.

8.2.1 Surface Analysis

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

A Readability

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

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

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

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

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

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

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

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

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

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

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

B Trins and Proxes

Page's initial study was centered on the idea that computers can be used to automatically evaluate and grade student essays using only statistically and easily detectable attributes, as effective as human teachers (Page 1966, 1968; Wresch 1993). In order to perform a statistical analysis, Page correlated two concepts: *proxes* (computer approximations of interest) with *human trins* (intrinsic variables – human measures used for evaluation) for better quantifying an essay's complexity. A correlation of *.71* proved that computer programs could predict grades quite reliably, similar to the inter-human correlation. Starting for Page's metrics of automatically grading essays and taking into consideration Slotnick's method (Slotnick 1972; Wresch 1993) of grouping proxes based on their intrinsic values, the following categories were used within our model for estimating textual complexity (see Table 25).

Table 25 *ReaderBench* (2) Surface analysis factors

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

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

C Entropy

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

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

8.2.2 Metrics for Word Complexity

From a different perspective, word complexity was treated as a combination of the following factors: syllable count, distance between the inflected form, lemma and

stem, whereas specificity is reflected in inverse document frequency from the training corpora, the distance in hypernym tree and the word polysemy count from the ontology. As an overview of the entire discourse, all these metrics are computed in a simple manner, by summing up the relevant values for all the words within text (only dictionary words after the initial NLP pipe processing) and then dividing the sum by the total number of words.

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

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

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

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

8.2.3 Morphology and Syntax

A Complexity, Accuracy and Fluency

Complexity, accuracy, and fluency (CAF) measures of texts have been used in linguistic development and in second language acquisition (SLA) research (House and Kuiken 2009). *Complexity* captures the characteristic of a learner's language, reflected in a wider range of vocabulary and grammatical constructions, as well as communicative functions and genres (Schulze 2010). *Accuracy* highlights a text's conformation to our experience with other texts, while *fluency*, in oral communication, captures the actual volume of text produced in a certain amount of time. Similar to the previous factors, these measures play an important role in automated essay scoring and textual complexity analysis. Schulze (2010)

considered that selected *complexity* measures should be divided into two main facets of textual complexity: sophistication (richness) and diversity (variability of forms). The defined measures depend on six units of analysis: letter (*l*), word form (*w*), bigram (*b* – groups of two words) and period unit (*p*), word form types (*t*) and unique bigrams (*u*). Additionally, textual complexity is devised into lexical and syntactic complexity:

Lexical Complexity:

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

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

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

$$v_2 = \frac{l}{w}, with\ 1 \leq v_2 \leq l \tag{33}$$

Syntactic Complexity:

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

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

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

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

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

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

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

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

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

B Part-of-Speech Statistics and Parsing Tree Structure

Starting from different linguistic categories of lexical items, our aim is to convert morphological information regarding the words and the sentence structure into relevant metrics to be assessed in order to better comprehend textual complexity. In this context, parsing and part of speech (POS) tagging play an important role in the morphological analysis of texts, in terms of textual complexity, by providing two possible vectors of evaluation: the normalized frequency of each part of speech and the structural factors derived from the parsing tree. Although the most common parts of speech used in discourse analysis are nouns and verbs, our focus was aimed at prepositions, adjectives and adverbs that dictate a more elaborate and complex structure of the text. Moreover, pronouns, that through their use indicate the presence of co-references, also indicate a more intertwined and complex structure of the discourse. On the other hand, multiple factors can be derived from analyzing the structure of the parsing tree: an increased number of leafs, a greater overall size of the tree and a higher maximum depth indicate a more complex structure, therefore an increased textual complexity (Gervasi and Ambriola 2002).

8.2.4 Semantics

Firstly, as seen in *2.1* Coherence and Comprehension, *textual complexity* is linked to *cohesion* in terms of comprehension; in other words, in order to understand a text, the reader must first create a well-connected representation of the information withheld, a situation model (van Dijk and Kintsch 1983) (see Figure 2). This connected representation is based on linking related pieces of textual information that occur throughout the text. Therefore, cohesion reflected in the strength of inner-block and inter-block links extracted from the cohesion graph influences readability, as semantic similarities govern the understanding of a text. In this context, discourse cohesion is evaluated at a macroscopic level as the average value of all links in the constructed cohesion graph (Dascalu et al. 2013a; Trausan-Matu et al. 2012a).

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

Thirdly, *entity-density features* proved to influence readability as the number of entities introduced within a text is correlated to the working memory of the text's targeted readers. In general, entities consisting of general nouns and named entities (e.g., people's names, locations, organizations) introduce conceptual information by identifying, in most cases, the background or the context of the text. More specifically, entities are defined as a union of named entities and general nouns (nouns and proper nouns) contained in a text, with overlapping general nouns removed. These entities have an important role in text comprehension due to the fact that established entities form basic components of concepts and propositions on which higher level discourse processing is based (Feng et al. 2010). Therefore, the entity-density factors focus on the following statistics: the number of entities (unique or not) per document or sentence, the percentages of named entities per document, the percentage of overlapping nouns removed or the percentage of remaining nouns in total entities.

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

8.2.5 Combining Textual Complexity Factors through Support Vector Machines

All the measures previously defined capture in some degree different properties of the analyzed text (readability, fluency, language diversity and sophistication, morphological structure, cohesion, etc.) and therefore can be viewed as attributes that describe the text. In order to use these attributes to estimate the complexity of the text, we have used a classifier that accepts as inputs text attributes and outputs the minimum grade level required by a reader to comprehend the specified text. In our integrated textual complexity analysis model we have opted for Support Vector Machine (SVM) classifiers that have been proven to be the most appropriate (François and Miltsakaki 2012; Petersen and Ostendorf 2009). A SVM (Cortes and Vapnik 1995; Press et al. 2007) is typically a binary linear classifier that maps the input texts seen as d-dimensional vectors to a higher dimensional space (hyperspace) through the mapping of a kernel function, in which, hopefully, these vectors are linearly separable by a hyperplane (see Figure 47).

Due to the fact that binary classifiers can map objects only into two disjoint classes, our multiclass problem can be solved using multiple Support Vector

Machines, each classifying a category of texts with different predefined classes of complexity (Duan and Keerthi 2005; Hsu and Lin 2002). 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).

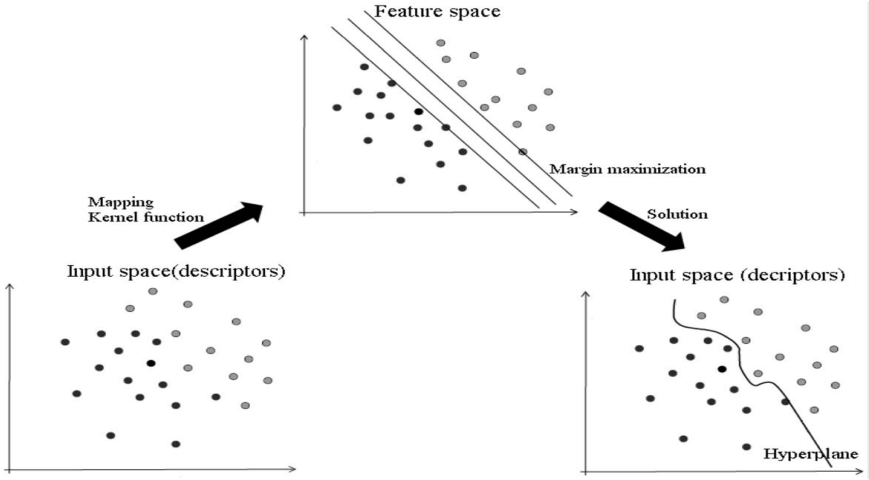


Fig. 47 General binary SVM mapping and separation through a hyperplane – adapted from Kozak et al. (2009)

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

8.2.6 Validation of the Integrated Textual Complexity Analysis Model

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

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

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

This validation scenario consisting of approximately 1,000 documents was twofold: we wanted, on one hand, to prove that the complete model is adequate and reliable and, on the other, to demonstrate that high level semantic features provide relevant insight that can be used for automatic classification. In the end, *k*-fold cross validation (Geisser 1993) was applied for extracting the following performance features (see Table 27 and Figure 48): precision or exact agreement (EA) and adjacent agreement (AA) (François and Miltsakaki 2012), as the percent to which the SVM was close to predicting the correct classification (Dascalu et al. in press; Dascalu et al. 2013a).

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

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

Table 27 *ReaderBench* (2) Textual complexity dimensions

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

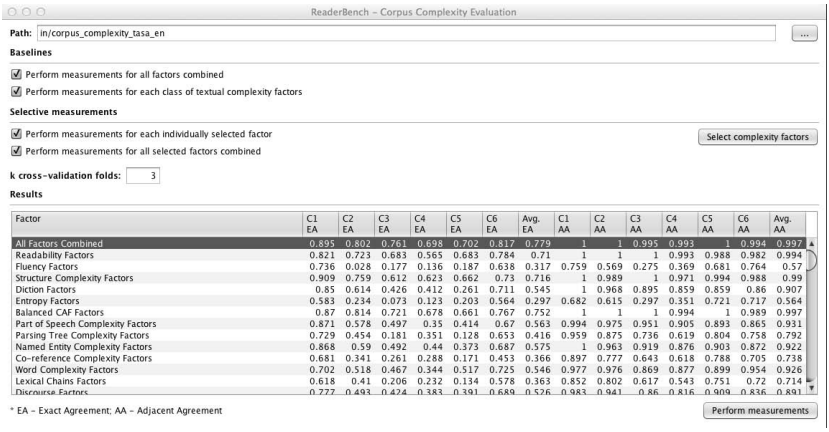
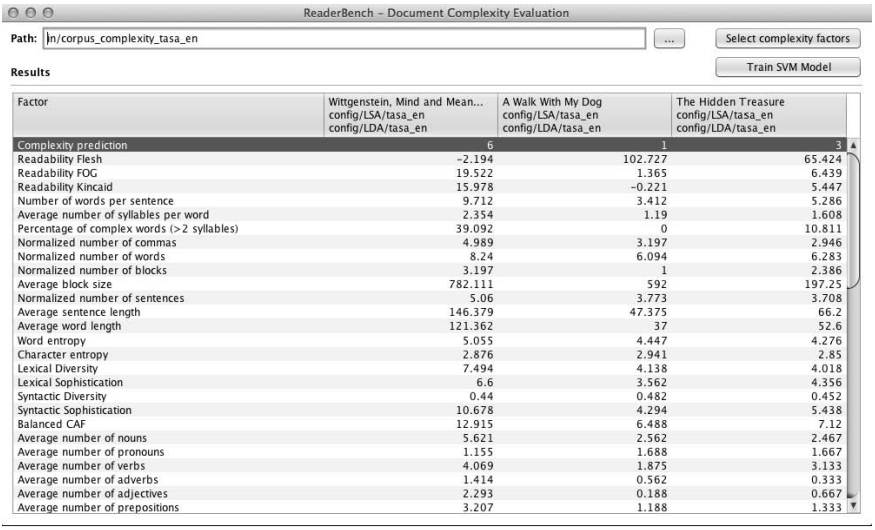


Fig. 48 *ReaderBench* (2) Textual complexity evaluation.Starting from a pre-processed corpus, the user has the opportunity to perform the following measurements applied on: 1/ the complete SVM model with all factors integrated; 2/ each individualcomplexity dimension (a predefined subset of textual complexity metrics); 3/ a specific set of selectedcomplexity factors, on which individual measurements or a single combined evaluation can beperformed. In the end, a table is automatically generated including the used factor (individual, textualcomplexity dimension or specific aggregation), exact and adjacent agreements for each complexity classfrom the corpus, as well as the average agreement values

Moreover, besides the factors presented in detail in Dascalu et al. (2012) that were focused on a more shallow approach, of particular interest was how semantic factors correlate to classic readability measures (Dascalu et al. 2013a). In this context, two additional measurements were performed. Firstly, an integration of all metrics from all textual complexity dimensions proved that the SVMs results are compatible with the DRP scores (EA = .779 and AA = .997), and that they provide significant improvements as they outperform any individual dimension precisions. The second measurement (EA = .597 and AA = .943) used only morphology and semantic measures in order to avoid a circular comparison between factors of similar complexity, as the DRP score is based on shallow factors. This result showed a link between low-level factors (also used in the DRP score) and in-depth analysis factors, which can also be used to accurately predict the complexity of a reading material (Dascalu et al. in press).

In terms of usability, besides the possibility to train and evaluate new textual complexity models on a given corpora (see Figure 48), *ReaderBench* enables tutors to assess the complexity of new reading materials based on the selected complexity factors and a pre-assessed corpus of texts, pertaining to different complexity dimensions. By comparing multiple loaded documents, tutors can better grasp each evaluation factor, refine the model to best suit their interests in terms of the targeted measurements and perform new predictions using only their features (see Figure 49).



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. 49 *ReaderBench* (2) Document complexity evaluation. Based on a pre-trained corpus, the user selects the complexity factors to be automatically used within the SVM model (by default all factors are pre-selected) and *ReaderBench* generates a complexity prediction for each loaded document, as well as all values corresponding to the selected individual factors in order to have a comparison of the evolution of specific metrics between different documents

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

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

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

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

Table 29 *ReaderBench* versus *Dmesure* (François and Miltsakaki 2012; François 2012)

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

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

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

References

Benjamin, R.G.: Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review 24, 63–88 (2012)

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

Brown, J.D.: An EFL readability index. JALT Journal 20(2), 7–36 (1998)

Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 27:21–27:27 (2011)

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

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C. (eds.): Introduction to Algorithms, 3rd edn. MIT Press, Cambridge (2009)

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

Dahl, R.: Matilda (trans: Robillot H). Folio Junior (Livre 744). Gallimard, Paris (2007)

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., Nardy, A.: Mining Texts, Learners Productions and Strategies with ReaderBench. In: Peña-Ayala, A. (ed.) Educational Data Mining: Applications and Trends. SCI. Springer (in press)

Dascalu, M., Dessus, P., Trausan-Matu, Ş., Bianco, M., Nardy, A.: ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS, vol. 7926, pp. 379–388. Springer, Heidelberg (2013)

Dascălu, M., Trausan-Matu, S., Dessus, P.: Towards an integrated approach for evaluating textual complexity for learning purposes. In: Popescu, E., Li, Q., Klamma, R., Leung, H., Specht, M. (eds.) ICWL 2012. LNCS, vol. 7558, pp. 268–278. Springer, Heidelberg (2012)

Davison, A., Kantor, R.: On the failure of readability formulas to define readable texts: A case study from adaptations. Reading Research Quarterly 17, 187–209 (1982)

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)

Dessus, P., Bianco, M., Nardy, A., Toffa, F., Dascalu, M., Trausan-Matu, S.: Automated analysis of pupils' self-explanations of a narrative text. In: de Vries, E., Scheiter, K. (eds.) Staging knowledge and experience, Meeting of the EARLI SIG 2 "Comprehension of Text and Graphics", Grenoble, France, pp. 52–54. LSE, Pierre-Mendès-France University (2012)

Duan, K.-B., Keerthi, S.S.: Which Is the Best Multiclass SVM Method? An Empirical Study. In: Oza, N.C., Polikar, R., Kittler, J., Roli, F. (eds.) MCS 2005. LNCS, vol. 3541, pp. 278–285. Springer, Heidelberg (2005)

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), Beijing, China, pp. 276–284. ACL (2010)

Flesch, R.: A new readability yardstick. Journal of Applied Psychology 32(3), 221–233 (1948)

François, T.: Les apports du traitement automatique du langage à la lisibilité du français langue étrangère. Université Catholique de Louvain, Faculté de Philosophie, Arts et Lettres, Louvain-la-Neuve, Belgium (2012)

François, T., Miltsakaki, E., Do, N.L.P.: machine learning improve traditional readability formulas? In: First Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2012), Montreal, Canada, pp. 49–57. ACL (2012)

Galley, M., McKeown, K.: Improving Word Sense Disambiguation in Lexical Chaining. In: Gottlob, G., Walsh, T. (eds.) 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), Acapulco, Mexico, pp. 1486–1488. Morgan Kaufmann Publishers, Inc. (2003)

Geisser, S.: Predictive inference: an introduction. Chapman and Hall, New York (1993)

Gervasi, V., Ambriola, V.: Quantitative assessment of textual complexity. In: Barbaresi, M.L. (ed.) Complexity in Language and Text, Plus, Pisa, Italy, pp. 197–228 (2002)

Graesser, A.C.: An introduction to strategic reading comprehension. In: McNamara, D.S. (ed.) Reading Comprehension Strategies: Theories, Intervention and Technologies, pp. 3–26. Erlbaum, Mahwah (2007)

Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behavioral Research Methods, Instruments, and Computers 36(2), 193–202 (2004)

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

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

Gunning, R.: The technique of clear writing. McGraw-Hill, New York (1952)

House, A., Kuiken, F.: Complexity, Accuracy, and Fluency in Second Language Acquisition. Applied Linguistics 30(4), 461–473 (2009)

Hsu, C.-W., Lin, C.-J.: A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks 13(2), 415–425 (2002)

Hsu, C.W., Chang, C.-C., Lin, C.-J.: A practical guide to support vector classification. National Taiwan University, Taipei (2010)

Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of New Readability Formulas (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Chief of Naval Technical Training, Naval Air Station Memphis (1975)

Kozak, K., Agrawal, A., Machuy, N., Csucs, G.: Data Mining Techniques in High Content Screening: A Survey. Journal of Computer Science & Systems Biology 2, 219–239 (2009)

Kukemelk, H., Mikk, J.: The Prognosticating Effectivity of Learning a Text in Physics. Quantitative Linguistics 14, 82–103 (1993)

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)

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)

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: Fifteenth Conference on Computational Natural Language Learning: Shared (TaskCONLL Shared Task 2011), Portland, OR, pp. 28–34. ACL (2011)

Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, vol. 1. Cambridge University Press, Cambridge (2008)

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

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

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, p. 27. R&L Education, Lanham (in press)

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

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 (2007b)

Millis, K., Magliano, J.P., Todaro, S.: Measuring discourse-level processes with verbal protocols and Latent Semantic Analysis. Scientific Studies of Reading 10(3), 225–240 (2006)

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

O'Reilly, T.P., Sinclair, G.P., McNamara, D.S.: iSTART: A Web-based Reading Strategy Intervention that Improves Students' Science Comprehension. In: Kinshuk, S.D.G., Isaías, P. (eds.) Int. Conf. Cognition and Exploratory Learning in Digital Age (CELDA 2004), Lisbon, Portugal, p. 8. IADIS Press (2004)

Oprescu, B., Dascalu, M., Trausan-Matu, S., Dessus, P., Bianco, M.: Automated Assessment of Paraphrases in Pupil's Self-Explanations. Scientific Bulletin, University Politehnica of Bucharest, Series C (in press)

Oprescu, B., Dascalu, M., Trausan-Matu, S., Rebedea, T., Dessus, P., Bianco, M.: Analiza automata a auto-explicatiilor. Revista Romana de Interactiune Om-Calculator 5(2), 71–76 (2012)

Page, E.: The imminence of grading essays by computer. Phi Delta Kappan 47, 238–243 (1966)

Page, E.: Analyzing student essays by computer. International Review of Education 14(2), 210–225 (1968)

Petersen, S.E., Ostendorf, M.: A machine learning approach to reading level assessment. Computer Speech and Language 23, 89–106 (2009)

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes: The Art of Scientific Computing, 3rd edn. Cambridge University Press, New York (2007)

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., Manning, C.A.: Multi-Pass Sieve for Coreference Resolution. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2010), Cambridge, MA, pp. 492–501. ACL (2010)

Sagot, B.: WordNet Libre du Francais (WOLF). INRIA, Paris (2008)

Sagot, B., Darja, F.: Building a free French wordnet from multilingual resources. In: Ontolex 2008, Marrakech, Maroc, p. 6 (2008)

Schmidt, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Int. Conf. on New Methods in Language Processing, Manchester, UK, vol. 4, pp. 44–49. Citeseer (1994)

Schmidt, H.: TreeTagger -a language independent part-of-speech tagger. Institute for Computational Linguistics, University of Stuttgart, Stuttgart (1995)

Schulze, M.: Measuring textual complexity in student writing. In: American Association of Applied Linguistics (AAAL 2010), Atlanta, GA, pp. 590–619. Waterloo Centre for German Studies (2010)

Shannon, C.E.: A Mathematical Theory of Communication. The Bell System Technical Journal 27, 379–423, 623–656 (1948)

Shannon, C.E.: Prediction and entropy of printed English. The Bell System Technical Journal 30, 50–64 (1951)

Slotnick, H.: Toward a theory of computer essay grading. Journal of Educational Measurement 9(4), 253–263 (1972)

Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: HLT-NAACL 2003, Edmonton, Canada, pp. 252–259. ACL (2003)

Trabasso, T., van den Broek, P.: Causal thinking and the representation of narrative events. Journal of Memory and Language 24, 612–630 (1985)

Trausan-Matu, S., Dascalu, M., Dessus, P.: Textual Complexity and Discourse Structure in Computer-Supported Collaborative Learning. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 352–357. Springer, Heidelberg (2012)

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

Vitale, M.R., Romance, N.R.: A knowledge-based framework for unifying content-area reading comprehension and reading comprehension strategies. In: McNamara, D.S. (ed.) Reading Comprehension Strategies, pp. 73–104. Erlbaum, Mahwah (2007)

Wolfe, M.B.W., Magliano, J.P., Larsen, B.: Causal and semantic relatedness in discourse understanding and representation. Discourse Processes 39(2-3), 165–187 (2005)

Wresch, W.: The imminence of grading essays by computer—25 years later. Computers and Composition 10(2), 45–58 (1993)