Biologically Inspired Cognitive Architectures (2014) 7, 39–52

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bica

RESEARCH ARTICLE

Effects of text essay quality on readers' working memory by a computational model

J. Ignacio Serrano [a,*], M. Dolores del Castillo [b], Ángel Iglesias [a]

[a] Bioengineering Group, Centro de Automática y Robótica, Consejo Superior de Investigaciones Científicas (CSIC), Spain
[b] Bioengineering Group, Consejo Superior de Investigaciones Científicas (CSIC), Spain

Received 16 May 2013; received in revised form 23 October 2013; accepted 23 October 2013

KEYWORDS
Computational modeling;
Reading;
Working memory;
Discourse quality

Abstract
Assessment of essay quality, also called essay scoring, is a task that has been always carried out by human graders. Graders are usually asked to give their scores according to several determined linguistic/semantic criteria. These criteria are related to lexical, syntactical, semantical and discourse features of the texts. In order to replace human graders, automated essay scoring systems make use of statistics on the latter features in order to quantify the quality of the essays. However, there is a subjective component within the evaluation of the text quality that cannot be measured by artificial scorers. Text essays are a form of natural language communication and therefore they cause effects on readers and their cognitive functions. In the work presented in this paper, the dynamic effects that a read text causes on the working memory of readers are studied by means of a connectionist model of memory during reading. Besides, the correlation of those effects with the essay quality scores and text linguistic features is also analyzed. The biologically inspired model of memory includes mechanisms for emulating bounded cognition, getting a little closer to the BICA Challenge achievement. The results obtained also prove how BICA models can feedback Neuroscience and Psychology, thus closing the interdisciplinary loop.

Introduction

Paradoxically, *Natural* Language is *artificially* described by a set of rules worldwide (Pinker, 2000). Human beings are commonly taught to properly use language by following that set of rules. This way, the quality of a language expression or passage can be measured by contrasting it with the corresponding normative description of the language. Thus,

* Corresponding author. Address: Centro de Automática y Robótica, Consejo Superior de Invesigaciones Científicas (CSIC), Ctra. Campo Real km 0.200, La Poveda, 28500 Arganda del Rey, Spain. Tel.: +34 918711900; fax: +34 918717050.
E-mail address: jignacio.serrano@csic.es (J.I. Serrano).

2212-683X/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bica.2013.10.002

several linguistic features belonging to different language levels (lexical, syntax, semantics, discourse, topic, etc.) are frequently used to characterize language units (Wang & Brown, 2007) and make comparisons with normative rules.

Nonetheless, Natural Language is a capacity developed as a product of evolution, acquired for the main purpose of communicating with other subjects (this is rather a matter for anthropologists (Duranti, 1997)) with the intention to cause some effects on their feelings, thoughts and ultimately their behavior (Pinker, 1995). From this point, language processing can be considered as a form of coding/decoding (emitter/receptor) of intentions and thoughts into phonemes and graphemes. Consequently, human beings have developed this processing ability and, as a mind ability, it requires cognitive processing and resources.

In turn, Natural Language is a dynamic entity in constant evolution (Christiansen & Kirby, 2003). This evolution of Natural Language has always favored the language use and structures that facilitate communication. There are several factors that can make communication easier, and one of them lies in decoding simplicity and requirements: the less cognitive processing and resources required for decoding the better the understanding. From this statement, a good quality coding implies a soft and easy decoding process (coding means here language structure and composition). Consequently, language quality can be measured in terms of cognitive effects and requirements during the understanding process.

In spite of the advances in the field of neurophysiological signal acquisition (EEG, fMRI, MEG, PET, etc.) (Démonet, 2005), the measure of dynamic cognitive load and effects during language processing is still a challenge nowadays. For this reason, this paper presents a computational model of dynamic memory — Cognitive Reading Indexing Model (CRIM) — that emulates the cognitive processing of human beings during reading. Computational modeling allows monitoring and measuring the use and capacity of the internal mechanism and resources of the model. Unlike biologically inspired related models such as the Cambrias et al.'s (Cambria, Mazzocco, & Hussain, 2013), which is focused on the static extraction of emotions and polarity that a piece of text contains. The model used in this paper is based on dynamic measurements of working memory usage and capacity during essay reading. These dynamic measures are confronted with the essay scores given by human graders in order to find a correlation between the text quality and the effects on cognitive performance during reading. It is worth noting that this work is not an attempt for a better automated essay scorer. Firstly, it is a step ahead in the development of mechanisms that emulate how perceived stimuli modulate our cognitive functions (bounded cognition, Gigerenzer & Selten, 2002), which is a primary target of the BICA challenge. Secondly, it is another proof of concept on how biologically inspired models can help to give insight into the cognitive processes of the human mind.

The next section presents the most important approaches to characterize language with quantitative measures at different linguistics levels, with the aim of capturing the subjective essence of human criteria and therefore replacing human graders with automated scorers. Cognitive effects of language quality comments different psychological evidence that confirms the influence of language structure and form on the cognitive processing of comprehension, and more concretely the role that working memory (WM from now on) plays in this process. In A computational model of dynamic working memory during reading, a computational working memory model for reading is described, showing the monitoring capabilities that it offers. Materials and empirical procedure presents the experimental design and procedure to test the correlation of the essay quality and memory effects, followed by the significant results obtained. Finally, some concluding remarks and future work are discussed.

Automated measuring of language quality

One of the controversial matters regarding essay grading is subjectivity, which is thought to cause the grade variation between different human graders (Carrell, 1995). Subjectivity has often been considered as an unfair factor by students being evaluated. In order to overcome this ''problem'' as well as to save the long time spent in the essay assessment (Mason & Grove-Stephenson, 2002), automated scorers came out as a fine alternative (Valenti, Neri, & Cucchiarelli, 2003). The fundamentals of such systems is the quantification, by means of observable linguistics features, of the intrinsic variables that human raters take subjectively into account (called *trins* Hearst, 2000). For instance, the number of words of a text would represent fluency; word length variation would correlate with diction; and number of relative pronouns and different parts of speech (POS) would be related to complexity of sentence syntax (Page, 1994).

The latter mentioned features belong to the lexical and syntax levels. Other computational essay scoring systems make use of features at the semantic level. Many of them produce a statistics-based semantic representation (Leacock, 2004) of the texts and compare it with the ideal essay or master text (Jerrams-Smith, Soh, & Callear, 2001). Other systems extract features regarding the discourse/rhetorical level by measuring semantic coherence between consecutive sentences or tracking topic shifts (Burstein, Leacock, & Swartz, 2001; Higgins & Burstein, 2006; Higgins, Burstein, & Attali, 2006).

Although all these artificial scoring systems work relatively fine for concrete domains, they carry some drawbacks. Most of them apply some kind of machine learning method, which is generally supervised and therefore needs training data (Valenti et al., 2003). In this case, training data is composed of texts annotated by human subjects. Thus, training data is costly to construct, difficult to find in turn, and it is still loaded of subjectivity. In addition, most of the grading systems are optimized and evaluated against scores given by human graders. This evaluation and optimization methodology makes artificial systems overfit the concrete human graders.

The primary aim of the creation of automated essay grading systems was the ''use of computers to increase the understanding of the textual features and cognitive skills involved in the creation and comprehension of written texts'' (Valenti et al., 2003). It seems that knowledge about the correlation between textual features has been enriched since the first automated scorers. However, the same enrichment has not occurred in the cognitive counterpart.

A step forward in this latter aspect would have led to a better quantification of the subjective features used by human scorers. Note that these subjective features are not related to human preferences or likings but to the subjective perception of the text, i.e. the effects produced in the cognitive processing of written language understanding. In this sense, this paper presents a study of the effects of language quality on the cognitive processing of reading comprehension, not a better alternative to the current computational approaches of automated essay scoring.

Cognitive effects of language quality

As said before, natural language must be decoded to the corresponding meaning in order to be understood. This decoding process implies the use of several cognitive resources and processes. The participation of such mechanisms is even more relevant if the language to be understood represents a passage composed of different consecutive utterances with a complete meaning (as an essay, for instance).

Among all the cognitive resources implied, working memory (WM) seems to be crucial (Carretti, Borella, Cornoldi, & Beni, 2009). Specually, working memory capacity has been strongly correlated to reading comprehension ability in the literature (Carretti, Cornoldi, Beni, & Romanò, 2005). It is not so related to the amount of concepts that could be retained but rather with the keeping/elimination of irrelevant concepts (Beni & Palladino, 2000). Consequently, a text that introduces or makes the reader infer more irrelevant concepts causes a more intensive use of WM and, therefore, it makes comprehension more difficult. Additionally, WM capacity has been proven to influence management of information that contradicts the predictions made by the inferences during reading (Otten & Berkum, 2009). Thus, a contradictory text will cause a costly processing of WM. Last but not least, WM contributes to the integration of meanings in the construction of the situation model (Calvo, 2005): the more diffuse a text, the harder the cognitive work carried out by working memory.

The latter evidence points out the influence of the language correctness, structure and style on the relation between reading comprehension and WM memory capacity, on the one hand, and reading comprehension and function, on the other hand. Besides, there exists evidence for the direct effects of text structure and style on comprehension too. Just in the late sixties, Frase (1969) stated that the order of the sentences absolutely influences what is kept in memory and, therefore, it is possible to program the memory inputs by rearranging the sentences. In this line, there are studies that prove the effects of the combination of short and long consecutive sentences on the memory recall after reading (Saito & Miyake, 2004). The whole organization of a text has been also proven to influence the late memory recall (Yussen et al., 1991). Even the text format (plain text vs. hypermedia) affects the language comprehension (Lee & Tedder, 2003).

In summary, a text that favors comprehension should not overload the WM capacity and function. Texts that make and let the reader infer and process causal relations, either by means of clues (McDaniel, Hines, & Guynn, 2002) or structural pauses (Sinclair, Healy, & Bourne, 1989), are better understood by both good and poor comprehenders, although the latter ones obtain benefits of such a feature in a higher degree. So, from the existent evidence it seems that text quality can be determined by measuring the cognitive load and resources required during the reading/comprehension process. At this point is where computational cognitive modeling can help.

A computational model of dynamic working memory during reading

The model used in this work is called CRIM (Cognitive Reading Indexing Model) (Serrano, del Castillo, & Iglesias, 2006). This model emulates and modulates memory processes (working memory and long-term memory) during reading (Fig. 1). The model operates over a semantic—linguistic knowledge previously acquired that is constructed as described by Algorithm 1:

Algorithm 1. Construction of the Long-Term Semantic/Linguistic Knowledge.

```
1:  LongTermWorkingMemory = empty
2:  FOR every text_i in the InputTextCollection
3:    APPLY Stoplist to text_i
4:    LEMMATIZE words of text_i
5:    FOR every sentence_k in text_i
6:      FOR every word_j in sentence_k
7:        IF word_j is not in LongTermWorkingMemory
8:          Add word_j to LongTermWorkingMemory
9:          Total_appearances_of_word_j = 0
10:       Total_appearances_of_word_j
          = Total_appearances_of_word_j + 1
11:       FOR other words word_h in sentence_k after
      word_j
12:         IF exists Connection(word_j, word_h)
13:           Connection(word_j, word_h) =
Connection(word_j, word_h)+1
14:           Connection(word_h, word_j) =
Connection(word_h, word_j)+1
15:         ELSE
16:           Add Connection(word_j, word_h) = 1
17:           Add Connection(word_h, word_j) = 1
18:  -- Normalize connection weights
19:  FOR every word_i in LongTermWorkingMemory
20:    FOR every Connection(word_i, word_j)
21:      ConnectionWeight(word_i, word_j) =
          Connection(word_i, word_j)/
Total_appearances_of_word_i
```

The knowledge acquired contains the semantic associations among concepts observed during the past experience of reading, simulating the contents of a kind of long-term working memory (Ericsson & Kintsch, 1995). This knowledge is represented by a weighted net of asymmetrically interconnected concepts, where the weights of the connections

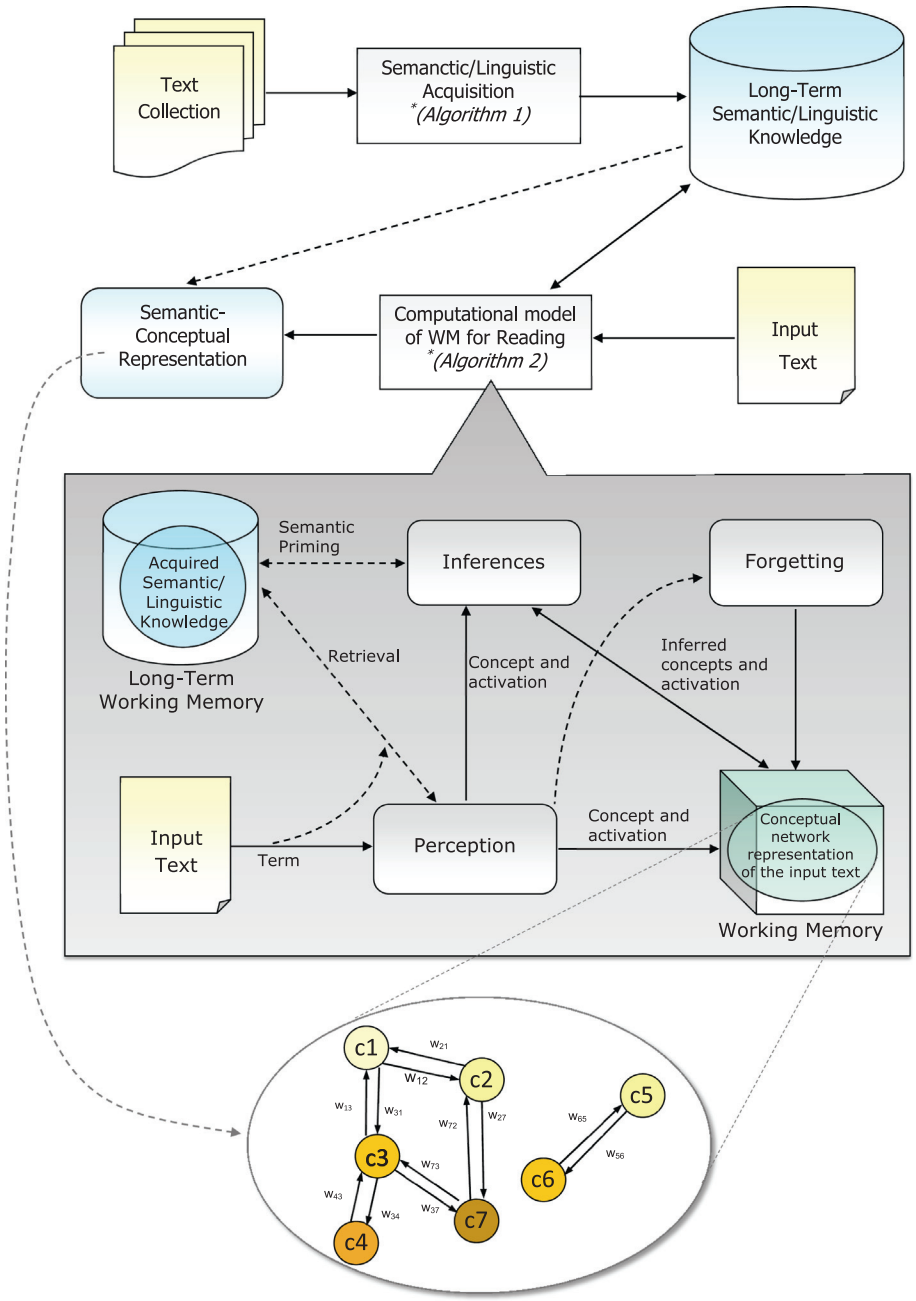


Fig. 1 Overall scheme of the dynamics of the computational model of working memory.

denote the semantic relatedness of the connected concepts. The semantic relatedness is given by the co-occurrence of concepts within the same sentence, so it is not structured by predefined relationships, such as AffectNet used in related works (Cambria et al., 2013; Cambria, Olsher, & Kwok, 2012).

Given the long-term semantic/linguistic knowledge, the model emulates reading of a natural language text and produces a semantic representation of it in WM (Fig. 1, middle). This representation is a net of associated concepts, each of them with a level of activation that indicates their signification within the text (Fig. 1, bottom). The model processes each word sequentially in the order they appear in the text. For each known word, i.e. present in the semantic knowledge, the corresponding concept is retrieved into working memory with a base level of activation. This level of activation recursively propagates to the associated concepts in long-term memory by decreasing the activation proportionally to the connection weights until reaching a minimum level, in a similar manner to Cambria et al. (2012). All the concepts reached by this propagation process are also retrieved into working memory with the propagated activation level. If a concept retrieved into working memory is already active on it, its activation on WM is increased by the retrieved activation. This way, each time a word is read a process of inference makes accessible other related concepts for future

prediction matching. Besides, the model always has a representation of the text at any time during reading. The following pseudocode describes the reading process:

Algorithm 2. Reading process of the computational model.

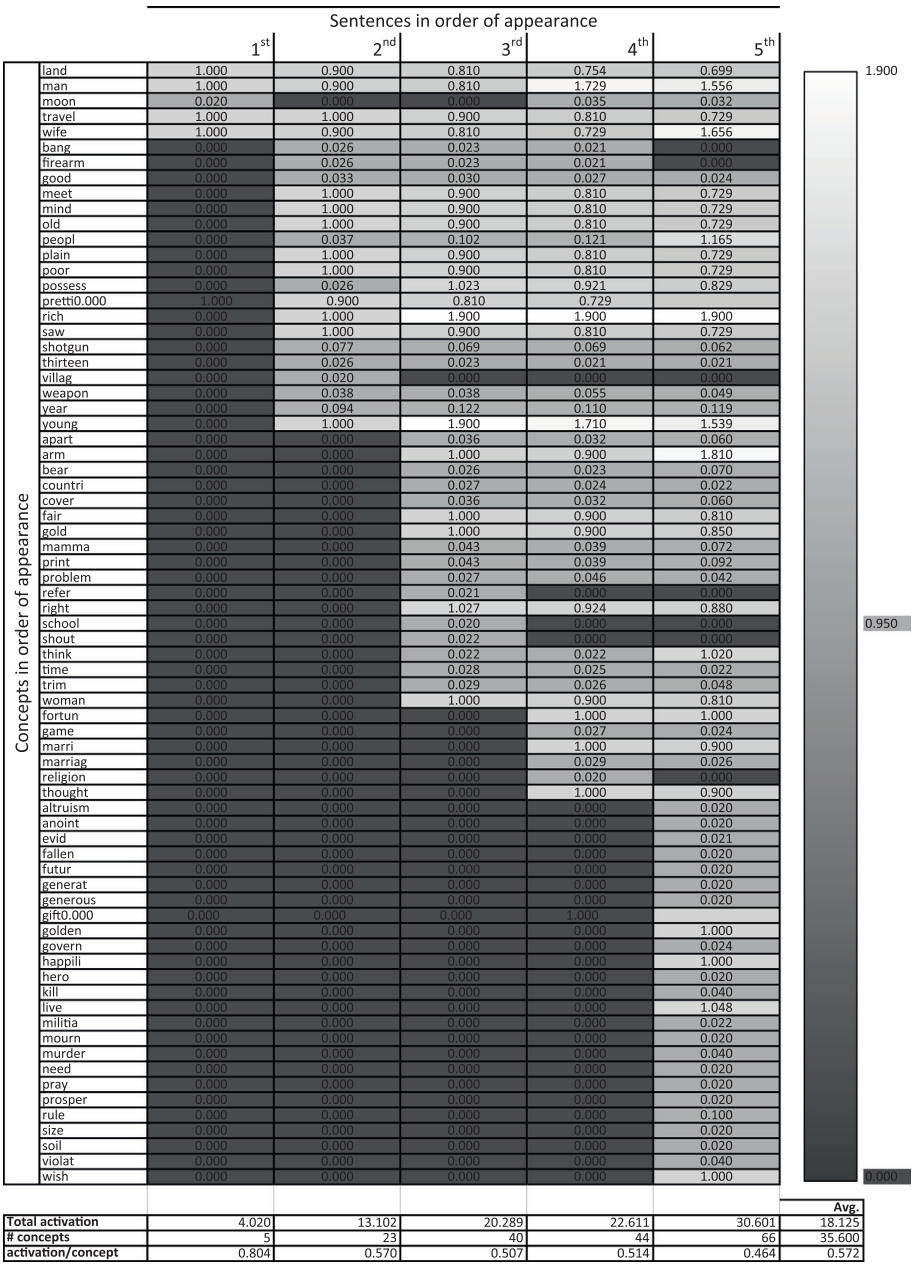
```
1:  WorkingMemory = empty
2:  FOR every sentence_k in InputText
3:    FOR every word_i in sentence_k
4:      IF word_i is in LongTermWorkingMemory
5:        Retrieve from LongTermWorkingMemory concept_i of word_i
6:        IF concept_i already in WorkingMemory
7:          Activation(concept_i) = Activation(concept_i) + BaseActivationLevel
8:        ELSE
9:          Add concept_i to WorkingMemory
10:         Activation(concept_i) = BaseActivationLevel
11:       --- Infer associated concepts from the one just read
12:       Spread(Activation(concept_i),concept_i, 1)
13:     --- Forget after each sentence. Decrease concept activation in WM
14:     IF END_OF_SENTENCE(InputText)
15:     FOR every concept_j in WorkingMemory
16:       Activation(concept_j) = Activation(concept_j)*ForgettingFactor
17:       IF Activation(concept_j) < PropagationThreshold
18:         Remove concept_j from WorkingMemory
19:   DEFINE Spread(activation, concept, level)
20:     IF activation < PropagationThreshold AND level < PropagationLevel
21:       FOR every neighbor_h of concept
22:         activationPropagated = activation *ConnectionWeight(concept, neighbor_h)
23:           IF neighbor_h already in WorkingMemory
24:             Activation(neighbor_h) = Activation(neighbor_h) + activationPropagated
25:           ELSE
26:             Add neighbor_h to WorkingMemory
27:             Activation(neighbor_h) = activationPropagated
28:           Spread(Activation(neighbor_h),neighbor_h, level+1)
```

Over the course of reading, the concepts stored in WM lose activation by the application of a decreasing factor at specific time intervals. In this case, time is controlled by the text structure itself because the decreasing factor is applied at the end of each sentence. Consequently, the concepts might be completely deleted from memory (forgotten) if they are not reactivated either by reading of the corresponding word or the propagation from another concept. In this sense, the model collects the influence of the discourse and style of the text on the comprehension. The model's behavior is thus modulated by different parameters such as forgetting factor, propagation threshold for both level and activation, base activation level, and WM capacity (in terms of number of concepts and total activation). These parameters refer to the retention, inference, attention and WM capacities, respectively. The setup of the parameters' values allows the model emulating the reading skills of singular individuals. In this work, the parameters were settled to emulate a skilled reader, matching the graders who score text essays. The parameters concerning WM capacity were ignored by setting them to infinite values. Since the target of the work is the study of the WM usage caused by the input text, no restrictions in WM capacity were imposed. All design aspects of the model have been created from psycholinguistics evidence in order to keep the model as plausible as possible (Serrano, del Castillo, & Iglesias, 2009b, 2009a).

WM load and activation of the model can be measured at any time, thus allowing the monitoring of the cognitive resources used. Fig. 2 shows the memory activation trace for all the concepts during the reading of the following text:

> *''There was once a man who traveled the land all over in search of a wife. He saw young and old, rich and poor, pretty and plain, and could not meet with one to his mind. At last he found a woman, young, fair, and rich, who possessed a right arm of solid gold. He married her at once, and thought no man so fortunate as he was. They lived happily together, but, though he wished people to think otherwise, he was fonder of the golden arm than of all his wife's gifts besides.''*

The model has been used in the past to correlate other factors such as prior knowledge, individual interest and reading engagement with memory and inference capacity (Serrano, del Castillo, & Iglesias, 2007; Serrano, del Castillo, & Iglesias, 2009a). In this paper, the target of the study is the relationship between text quality and style and dynamic working memory usage. Particularly, text quality is determined by human raters; memory usage is characterized by



Concepts in order of appearance	Sentences in order of appearance 1st	2nd	3rd	4th	5th
land	1.000	0.900	0.810	0.754	0.699
man	1.000	0.900	0.810	1.729	1.556
moon	0.020	0.000	0.000	0.035	0.032
travel	1.000	1.000	0.900	0.810	0.729
wife	1.000	0.900	0.810	0.729	1.656
bang	0.000	0.026	0.023	0.021	0.000
firearm	0.000	0.026	0.023	0.021	0.000
good	0.000	0.033	0.030	0.027	0.024
meet	0.000	1.000	0.900	0.810	0.729
mind	0.000	1.000	0.900	0.810	0.729
old	0.000	1.000	0.900	0.810	0.729
peopl	0.000	0.037	0.102	0.121	1.165
plain	0.000	1.000	0.900	0.810	0.729
poor	0.000	1.000	0.900	0.810	0.729
possess	0.000	0.026	1.023	0.921	0.829
pretti0.000	1.000	0.900	0.810	0.729	
rich	0.000	1.000	1.900	1.900	1.900
saw	0.000	1.000	0.900	0.810	0.729
shotgun	0.000	0.077	0.069	0.069	0.062
thirteen	0.000	0.026	0.023	0.021	0.021
villag	0.000	0.020	0.000	0.000	0.000
weapon	0.000	0.038	0.038	0.055	0.049
year	0.000	0.094	0.122	0.110	0.119
young	0.000	1.000	1.900	1.710	1.539
apart	0.000	0.000	0.036	0.032	0.060
arm	0.000	0.000	1.000	0.900	1.810
bear	0.000	0.000	0.026	0.023	0.070
countri	0.000	0.000	0.027	0.024	0.022
cover	0.000	0.000	0.036	0.032	0.060
fair	0.000	0.000	1.000	0.900	0.810
gold	0.000	0.000	1.000	0.900	0.850
mamma	0.000	0.000	0.043	0.039	0.072
print	0.000	0.000	0.043	0.039	0.092
problem	0.000	0.000	0.027	0.046	0.042
refer	0.000	0.000	0.021	0.000	0.000
right	0.000	0.000	1.027	0.924	0.880
school	0.000	0.000	0.020	0.000	0.000
shout	0.000	0.000	0.022	0.000	0.000
think	0.000	0.000	0.022	0.022	1.020
time	0.000	0.000	0.028	0.025	0.022
trim	0.000	0.000	0.029	0.026	0.048
woman	0.000	0.000	1.000	0.900	0.810
fortun	0.000	0.000	0.000	1.000	1.000
game	0.000	0.000	0.000	0.027	0.024
marri	0.000	0.000	0.000	1.000	0.900
marriag	0.000	0.000	0.000	0.029	0.026
religion	0.000	0.000	0.000	0.020	0.000
thought	0.000	0.000	0.000	1.000	0.900
altruism	0.000	0.000	0.000	0.000	0.020
anoint	0.000	0.000	0.000	0.000	0.020
evid	0.000	0.000	0.000	0.000	0.021
fallen	0.000	0.000	0.000	0.000	0.020
futur	0.000	0.000	0.000	0.000	0.020
generat	0.000	0.000	0.000	0.000	0.020
generous	0.000	0.000	0.000	0.000	0.020
gift0.000	0.000	0.000	0.000	1.000	
golden	0.000	0.000	0.000	0.000	1.000
govern	0.000	0.000	0.000	0.000	0.024
happili	0.000	0.000	0.000	0.000	1.000
hero	0.000	0.000	0.000	0.000	0.020
kill	0.000	0.000	0.000	0.000	0.040
live	0.000	0.000	0.000	0.000	1.048
militia	0.000	0.000	0.000	0.000	0.022
mourn	0.000	0.000	0.000	0.000	0.020
murder	0.000	0.000	0.000	0.000	0.040
need	0.000	0.000	0.000	0.000	0.020
pray	0.000	0.000	0.000	0.000	0.020
prosper	0.000	0.000	0.000	0.000	0.020
rule	0.000	0.000	0.000	0.000	0.100
size	0.000	0.000	0.000	0.000	0.020
soil	0.000	0.000	0.000	0.000	0.020
violat	0.000	0.000	0.000	0.000	0.040
wish	0.000	0.000	0.000	0.000	1.000

						Avg.
Total activation	4.020	13.102	20.289	22.611	30.601	18.125
# concepts	5	23	40	44	66	35.600
activation/concept	0.804	0.570	0.507	0.514	0.464	0.572

Fig. 2 Memory activation trace for all the concepts read and inferred in an example text. Activation levels (color intensity) for the different concepts (*y*-axis) at the end of each sentence (*x*-axis) are presented.

the factors described next, which represents a summarization of the memory activation trace. Notice that all factors presented below rely on the activation values of the concepts in WM at the end of each sentence of the read text (Fig. 2, bottom table):

- Mean Activation per Concept (MAC): At the end of each sentence the activation per concept is calculated as the mean activation of the concepts currently in WM. The MAC value is the mean of the activation per concept across all sentences.

$$MAC = \frac{\sum_{n\ sentences}^{i} \frac{total\ activation\ of\ WM\ after\ sentence\ i}{concepts\ in\ WM\ after\ sentence\ i}}{n}$$

being n the number of sentences of the text.

- Mean Deviation per Concept (MDC): In addition to the mean activation of the concepts, the standard deviation of that mean is also calculated at the end of each sentence. The MDC value is the mean of the standard deviations per concept across all sentences.

$$MDC = \frac{\sum_{n\ sentences}^{i} Std(\frac{total\ activation\ of\ WM\ after\ sentence\ i}{concepts\ in\ WM\ after\ sentence\ i})}{n}$$

being *Std* the Standard Deviation.

- Mean Absolute Difference (MAD): It stands for the difference of MAC value minus MDC value.

$$MAD = MAC - MDC$$

- Mean Relative Difference (MRD): It is calculated as MAD divided by the maximum among MAC and MDC.

$$MRD = \frac{MAD}{max(MAC, MDC)}$$

- Average Activation (AA): It refers to the average of the current WM activation values at the end of each sentence.

$$AA = \frac{\sum_{n\ sentences}^{i} total\ activation\ of\ WM\ after\ sentence\ i}{n}$$

- Maximum Activation (MA): It is the maximum activation value of WM among all sentences.

$$MA = max_{n\ sentences}^{i}(total\ activation\ of\ WM\ after\ sentence\ i)$$

- Average Concepts (AC): It stands for the average number of concepts of WM across all sentences.

$$AC = \frac{\sum_{n\ sentences}^{i} number\ of\ concepts\ in\ WM\ after\ sentence\ i}{n}$$

- Maximum Concepts (MC): It quantifies the maximum number of concepts in WM among all sentences.

$$MC = max_{n\ sentences}^{i}(number\ of\ concepts\ in\ WM\ after\ sentence\ i)$$

- Activation Difference (AD): It just represents the difference of MA minus AA.

$$AD = MA - AA$$

- Concept Difference (CD): It is the difference of MC minus AC.

$$CD = MC - AC$$

- Final Mean Activation per Concept (fMAC): Mean activation of the concepts currently in WM at the end of the text (analogous to MAC).
- Final Mean Deviation per Concept (fMDC): In addition to the final mean activation of the concepts, the standard deviation of that mean is also calculated at the end of the text (analogous to MDC).
- Final Mean Absolute Difference (fMAD): It stands for the difference of fMAC value minus fMDC value.

$$fMAD = fMAC - fMDC$$

- Final Mean Relative Difference (fMRD): It is calculated as fMAD divided by the maximum among fMAC and fMDC.

$$fMRD = \frac{fMAD}{max(fMAC, fMDC)}$$

- Reading Time (RT): Time in milliseconds spent by the model in reading the text.

Materials and empirical procedure

For the construction of the linguistic knowledge of the model, the well-known Reuters-21578 Text Collection Data Set was used. It is composed of 21,578 texts in English from the Reuters newswire on 1987. The documents were processed and made available by Reuters Ltd. and Carnegie Group, Inc. It contains about 218 million words (actually tokens) in total.

For the study, the text material used consisted of three collections of essays: firstly, nine sample essays from College Board web site,[1] each of them scored by two experienced high school teachers (CB collection from now on), with mean scores 6, 6, 5, 5, 4, 4, 2, 3, 1, respectively; secondly, seven essays of Hunter College web site,[2] each of them scored by two human graders with mean scores of 6, 5, 4, 3, 2, 1, 1, respectively (denoted as HC collection); finally, seven essays of Dr. Li's Secret web site,[3] each of them scored by one experienced teacher with scores 6, 5.5, 5, 5, 3.5, 3, 2.5, respectively (ME collection from now on). The given scores are on the 6–1 scale, being 6 the highest and 1 the lowest. All the essays were written by students applying for SAT (Scholastic Assessment Test). Within each of the three collections, the topic of the essays remains constant.

The model is modulated by certain parameters of memory capacity, inference depth and concept retention that determine the reading skills or ability. Consequently, these parameters were settled to the best possible values in order to match the degree of expertise of the human graders (Serrano et al., 2009b, 2009a).

Each text was then input into the model for reading, and for each text reading all the factors described in the previous section were measured and stored. Then, a correlation analysis between those factors and the scores of the texts was carried out.

Results

After applying the model to the input texts, the factor values (columns) from the reading of each of them (rows) were obtained and presented in Table 1.

Next, the Pearson correlation coefficients r between each factor (columns) and the score are presented in Table 2, either for each of the three essays collections and for all the essays together (rows).

The results show a direct correlation between the scores and the factors related to total values of both activation (AA, MA) and number of concepts (AC, MC) in WM, although it is only statistically significant for AA in the three collections of essays. Consequently, the higher scored text the more activation and concepts in WM. This implies that a high scored text contains related concepts that appear frequently in consecutive sentences, which makes the activation increase, and that there are slight changes to related topics along the text, which makes the number of concepts grow while keeping the previous ones active in WM. From here, it is derived that text quality can be determined to some extent in terms of several factors concerning the dynamic use and state of WM (the activation rather than the number of items) during reading.

Although the tendencies of the results are similar for the three collections of texts, the correlation values are higher

[1] http://www.collegeboard.com/student/testing/sat/prep_one/essay/pracStart.html.
[2] http://rwc.hunter.cuny.edu/reading-writing/on-line/scoring-and-sample-essays.html.
[3] http://www.mathenglish.com/Program/Essay/Info/DrLi-writing-Scoring-Samples.htm.

Table 1 Dynamic factor values measured from the reading of the input texts by the computational model (CB = College Board, HC = Hunter College, ME = Mathenglish).

Source	Score	MAC	MDC	MAD	MRD	AA	MA	AC	MC	AD	CD	fMAC	fMDC	fMAD	fMRD	RT
CB	1	0.055	0.260	−0.205	−0.805	10.43	19.89	89.00	168	9.47	79.00	0.098	0.350	−0.252	−0.720	0.110
CB	2	0.141	0.342	−0.200	−0.607	21.73	35.86	104.00	127	14.13	23.00	0.244	0.481	−0.237	−0.492	0.078
CB	3	0.077	0.279	−0.202	−0.759	40.15	79.73	276.25	511	39.57	234.75	0.159	0.455	−0.296	−0.650	0.760
CB	4	0.080	0.246	−0.166	−0.696	29.63	45.24	157.50	240	15.61	82.50	0.119	0.319	−0.200	−0.627	0.211
CB	4	0.068	0.254	−0.187	−0.752	43.91	68.63	306.21	454	24.72	147.79	0.090	0.297	−0.207	−0.698	0.780
CB	5	0.075	0.250	−0.175	−0.711	50.12	69.59	308.00	477	19.48	169.00	0.110	0.334	−0.224	−0.671	0.738
CB	5	0.064	0.238	−0.174	−0.764	35.42	62.42	207.83	407	27.00	199.17	0.118	0.337	−0.219	−0.649	0.561
CB	6	0.070	0.227	−0.157	−0.722	36.58	69.69	233.57	448	33.10	214.43	0.131	0.347	−0.216	−0.622	0.392
CB	6	0.073	0.244	−0.171	−0.714	36.25	52.63	205.00	300	16.38	95.00	0.099	0.285	−0.185	−0.652	0.346
HC	1	0.062	0.257	−0.195	−0.788	16.88	31.11	148.43	279	14.23	130.57	0.116	0.414	−0.297	−0.719	0.255
HC	1	0.079	0.298	−0.219	−0.771	16.53	31.75	134.60	222	15.22	87.40	0.148	0.442	−0.294	−0.665	0.184
HC	2	0.082	0.290	−0.207	−0.746	20.63	35.60	134.73	228	14.97	93.27	0.150	0.438	−0.287	−0.656	0.222
HC	3	0.072	0.251	−0.179	−0.743	43.52	78.84	358.44	565	35.32	206.56	0.119	0.341	−0.221	−0.650	1.263
HC	4	0.110	0.330	−0.220	−0.708	87.70	161.88	439.25	757	74.18	317.75	0.171	0.397	−0.226	−0.570	3.369
HC	5	0.095	0.292	−0.197	−0.718	69.85	152.06	431.53	669	82.22	237.47	0.186	0.436	−0.250	−0.572	3.455
HC	6	0.107	0.304	−0.197	−0.695	94.61	165.99	533.18	807	71.38	273.82	0.211	0.459	−0.248	−0.540	6.940
ME	2.5	0.084	0.272	−0.188	−0.703	34.67	47.85	232.00	350	13.18	118.00	0.112	0.320	−0.208	−0.649	0.489
ME	3.5	0.128	0.364	−0.235	−0.683	23.78	46.22	94.47	164	22.44	69.53	0.213	0.485	−0.272	−0.560	0.197
ME	3	0.056	0.210	−0.153	−0.741	30.17	50.23	246.85	407	20.06	160.15	0.098	0.294	−0.196	−0.666	0.494
ME	5.5	0.065	0.227	−0.161	−0.724	46.83	89.86	285.52	515	43.03	229.48	0.119	0.339	−0.220	−0.649	1.037
ME	5	0.074	0.250	−0.176	−0.723	56.29	91.94	382.36	526	35.65	143.64	0.130	0.381	−0.252	−0.660	1.705
ME	5	0.062	0.227	−0.165	−0.743	52.18	98.63	330.26	607	46.45	276.74	0.115	0.328	−0.213	−0.649	1.457
ME	6	0.069	0.267	−0.197	−0.760	53.08	88.34	317.17	538	35.27	220.83	0.087	0.310	−0.224	−0.721	1.129

Table 2 Pearson correlation coefficients between the text scores and each factor values from the model (CB = College Board, HC = Hunter College, ME = Mathenglish).

Source	r_{MAC}	r_{MDC}	r_{fMAD}	r_{fMRD}	r_{AA}	r_{MA}	r_{AC}	r_{MC}	r_{AD}	r_{CD}	r_{fMAC}	r_{fMDC}	r_{fMAD}	r_{fMRD}	r_{RT}
CB	−0.291	−0.668*	0.888*	0.051	0.705*	0.620	0.572	0.561	0.364	0.469	−0.382	−0.569	0.619	−0.079	0.395
HC	0.807*	0.426	0.185	0.946*	0.939*	0.945*	0.955*	0.928*	0.924*	0.841*	0.840*	0.169	0.697	0.944*	0.941*
ME	−0.338	−0.235	0.137	−0.578	0.822*	0.904*	0.613	0.707	0.884*	0.688	−0.267	−0.107	−0.156	−0.467	0.742
All	−0.054	−0.288	0.464*	0.235	0.659*	0.606*	0.603*	0.611*	0.517*	0.557*	−0.098	−0.347	0.543*	0.161	0.437*

* $p < .05$.

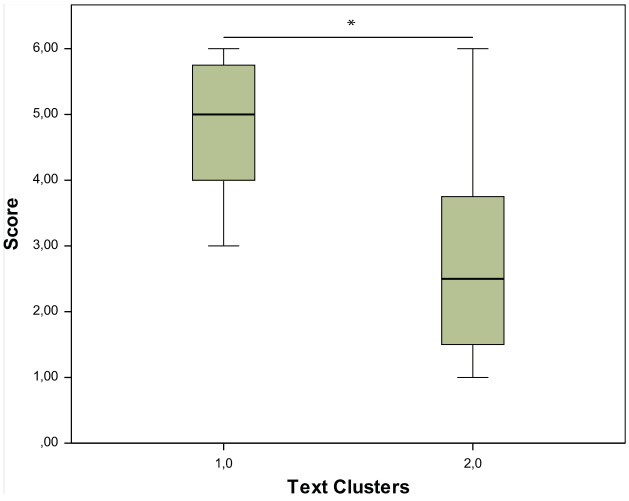


Fig. 3 Box diagram of the scores of the two clusters of texts obtained by the k-median algorithm (* $p < .01$).

for the Hunter College collection. Besides, there also exists a difference in the sign of the correlation concerning the factors related to the activation per concept (MAC, MDC, fMAC, fMDC) in CB and ME collection with respect to HC collection.

For a more detailed characterization of the text scores in terms of the model's factors, a clustering algorithm (k-median) was applied to the dataset. The six most globally correlated factors, which also presented significant correlation in one of the three collections at least (i.e. AA, MA, AC, MC, AD, CD), were selected as the input features to the clustering algorithm. The target of the algorithm was to find two clusters of texts that were significantly separated in the space formed by the input features. The results are shown in Fig. 3 and Table 3. This table presents average values of the centroids corresponding to the two clusters. In the last two rows, the average (standard deviation) and median scores of the texts belonging to each of the two clusters are shown.

The values obtained indicate that the algorithm grouped the texts by score. Cluster 1 represents the texts of higher scores, with a median score of 5.00, while Cluster 2 represents the texts with lower scores, with a median value of 2.00. All the model factors considered are significantly higher for Cluster 1. Therefore, the results show that good quality is related with intensive use of WM, in terms of activation and number of concepts. Intensive use here does not mean high load and effort, as it is discussed below.

Besides, the time that the model takes to read each text (RT) also seems to be directly correlated to the scores assigned (only statistically significant for the Hunter College collection). This might be due to several features of the texts, such as the number of words, the frequency of central words or the length of the sentences. These and other linguistic properties of the scored essays can be the source of the correlations found. In order to get deep into this question, Pearson correlation coefficients between diverse linguistic properties of the texts and the WM model factors are presented in Table 4 (only statistically significant values at $p < .05$ are presented for the sake of simplicity).

Table 3 Average values of the centroids of the two clusters of texts obtained by the k-median algorithm. Differences in all the rows are statistically significant (U Mann-Whitney, $^{*}p < .05$).

	Cluster 1	Cluster 2
AA	56.00	25.00
MA	101.00	41.00
AC	350.00	159.00
MC	572.00	262.00
AD	45.03	16.00
CD	222.00	103.00
Avg. score (std.)	4.00 (1.08)	2.00 (1.00)
Median score	5.00	2.00

The linguistic properties were extracted with the Coh-Metrix 2.0 tool (Graesser, McNamara, Louwerse, & Cai, 2004). Coh-Metrix has been developed at the Department of Psychology, University of Memphis. This tool is based in the concept of cohesion that the authors explicitly define as ''characteristics of the explicit text that play some role in helping the reader mentally connect ideas in the text'' (Graesser, McNamara, & Louwerse, 2004). It provides up to fifty four indexes representing properties of the texts concerning the mentioned concept of cohesion, ranging from lexical to discourse levels. From all output indexes, only the Coh-Metrix indexes that presented statistically significant correlation coefficients with the WM model factors are shown in Table 4 (Coh-Metrix features are named as their original identifiers).

The results can be divided in four groups as follows. First, the text features that are directly correlated with factors related to the average and maximum activation and number of concepts (AA, MA, AC, MC, AD, CD), and reading time (RT) of the WM model. The text features that present this kind of correlation are *DENCONDi*, *DENNEGi*, *CONi* and *SYNHw*. *DENCONDi*, refers to the number of conditional expressions per 1000 words; *DENNEGi*, refers to the number of negations per 1000 words; *CONi* denotes the number of connectives per 1000 words. Finally, *SYNHw* represents the mean number of higher level constituents (noun phrases and verb phrases) per word, so that it provides a measure of the structural (syntactic) density of the sentences. This way, a high number of conditional, negative expressions and connectives as well as a higher syntactic density make the mean activation and number of concepts increase in WM, as well as reading time spent by the model.

Second, the text features that correlate inversely with the model factors concerning the mean and maximum activation and number of concepts in WM, as well as reading time, are *HYNOUNaw*, *SYNNP* and *WORDCacw*. *HYNOUNaw* refers to the mean number of hypernyms (in Wordnet taxonomy) per noun; *SYNNP* is the mean number of modifiers per noun phrase (articles, adjectives, quantifiers, etc.); *WORDCacw* denotes the mean value of concreteness (rating from the MRC Psycholinguistics Database Wilson, 1988) of content words (nouns, adverbs, adjectives, main verbs, and other categories with rich conceptual content). The three mentioned features denote the same one way or another, that is, a kind of measure of how concrete or abstract the words in the text are (notice that the more modifiers has a noun the more concrete it becomes). Therefore, the more concrete words the less activation and number of concepts in WM and the shorter time spent in reading.

Third, another group of results can be formed by the text features that positively correlate with the model factors related to the mean activation and standard deviation per concept in WM (MAC, MAD, MRD, fMAD, fMRD). These features are *CONLGni* and *CONADni*. *CONLGni* refers to the number of negative logical connectives (such as ''*nor*'' or ''*neither*'') per 1000 words; *CONADni* denotes the number of negative additive connectives (such as ''*however*'' or ''*but*'') per 1000 words. Thus, the higher number of negative connectives the more activated the concepts in WM and the higher relative difference between the mean activation per concept and the standard deviation of such mean or, in other words, a more heterogeneous distribution of the activation of the concepts in WM.

Fourth, a final group of correlation results that can be mentioned is the one corresponding to the text feature that

Table 4 Statistically significant Pearson correlation coefficients between the Coh-Metrix indexes of texts and each factor values from the model of the graders ($p < .05$).

Coh-Metrix Concept	MAC	MAD	MRD	AA	MA	AC	MC	AD	CD	fMAC	fMRD	RT
CAUSVP						0.436						
CONTPpi									0.457			
CONADni			0.428									
CONi					0.438	0.432		0.443				0.475
DENCONDi				0.562	0.550	0.613	0.571	0.507	0.445			0.714
DENNEGi				0.505	0.437	0.532	0.458					0.681
DENLOGi												0.571
HYNOUNaw					−0.486			−0.562				−0.550
SYNNP				−0.446	−0.456			−0.442				−0.470
SYNHw				0.431	0.457			0.460				
TYPTOKc		0.441										
WORDCacw				−0.423	−0.438			−0.429				−0.574
CONLGpi							0.417					0.471
CONLGni	0.436		0.482							0.418	0.426	

is negatively correlated with the factors related to the mean activation per concept in WM. That feature is *TYPTOKc*, and it refers to the mean number of times that any word occurs within the text. Consequently, it can be viewed as a measure of the incidence of repetition of the words. In this case, the higher incidence of repetition of the words (corresponding to lower values of *TYPTOKc*) the higher difference between the mean activation per concept and the standard deviation. Therefore, the repetitive occurrences of the words produce a more heterogeneous distribution of the activation of the concepts in WM.

From the correlation results presented, it is derived that all the WM factors are modulated by one or more features of the texts belonging to different linguistics levels.

Discussion

The main effect of text quality on WM found is an increasing of the WM average activation and number of concepts with increasing scores. This might seem contradictory with the argument stated in the introduction that a good text should facilitate the cognitive process of comprehension by an optimum usage of WM. However, cognitive cost is mainly due to retrieval of concepts from long-term memory and inference processes. Concepts already present in WM are not needed to be retrieved from long-term memory when their reference term is perceived again or when they are inferred from a perceived associated concept. In addition, the more activated corresponding concepts in WM, the easier concept mapping during the perception process as well as the inference making. Consequently, if WM contains and keeps most of the concepts that will appear or be inferred along the text with a high activation, the comprehension process (perception, integration, inference) will be more straight and easier. This is what the correlation found actually means. Besides, there is evidence that confirms that keeping a high number of highly activated concepts in WM requires just attention and engagement (Awh, Vodel, & Oh, 2006). Therefore, extensive use and activation of WM do not mean here a high cognitive effort. In fact, they correlate with an optimum WM usage. It is also worth noting that high activation and usage are relative to the texts used, and they do not necessarily mean absolute high values.

In the result section, a certain variability of results among the three collections has been found. This divergence might be due to the differences among the distributions of scores in the three collections (Wang & Brown, 2007), being the Hunter College distribution the most homogeneous. These differences could also explain the inverse correlation between scores and factors related to the mean activation per concept, either along the text or at the end of it (MAC, MDC, fMAC, fMDC), both in College Board and in Mathenglish collections in contrast with Hunter College collection. The mean scores for CB and ME collections are 4.00(±1.73) and 4.36(±1.35), respectively. The mean score for HC collection is 3.14(±1.95). Consequently, in addition to the uniformity of the distributions, the mean quality of the studied essays must be a determinant factor. This way, it could be hypothesized that there exists a criterion for a fine-grained assessment of the quality of the best scored essays. So, among good essays, the best scored ones produce a more uniform and homogeneous distribution of the activation of concepts in working memory. This uniform and homogeneous distribution means a low average (and standard deviation) activation of the concepts. This might happen either because of long sentences or because the majority of the concepts in WM have been inferred and not directly perceived from the text. Since no correlation between WM usage and sentence length was found, the second argument is the most plausible. Nonetheless, the variability among collections could also be caused by the different human graders that scored each of them. Further research is needed to determine the source of this variability.

Among all correlation results, two relations are specially interesting to discuss. The first one concerns with the direct correlation between the assigned score and the difference between the average and maximum activation, and the average and maximum number of concepts. That is to say, the difference grows as the score increases. Consequently, a good text would induce a maximum level of activation and number of concepts at some instant that is much higher than the levels achieved during the rest of the text. It would be interesting to study where that maximum is reached for such high-scored texts. It could be hypothesized that the maximum is reached either at the beginning, where most concepts would be introduced and then sequentially elaborated along the text, or at the end, where most treated concepts would be summarized.

The second interesting evidence comes from the negative correlation between the repetition incidence of words (*TYPTOKc*, note that a high value of this parameter means a low repetition incidence) and the average activation difference per concept (MAD), i.e. the difference between the mean activation per concept and its corresponding standard deviation. If the words are not repeated along the text, their associated concepts will keep at decreasing low levels of activation, in contrast with the high level of activation of concepts corresponding to frequent and function words. However, this effect requires not only a low incidence of repetition, but also the presence of non-related words. If the words are somehow related, the presence of some of them would trigger the inference of the others, thus compensating the effects of the lack of repetition. So, the use of a varied vocabulary is profitable when the terms are semantically related. Otherwise, a sparse distribution of concepts in working memory is induced, which in turn hinders retrieval and inference and, hence, the reading process.

No correlation concerning the number of words, the number of sentences, the average size of the sentences, or any other statistic related to the text size has been found. Most automated essay scorers and measures of text quality are highly influenced by the size of the texts or sentences. However, these features have no significant influence in the WM usage, although it is also true that most factors measured from the model are average values.

Finally, it must be said that the same model configuration has been used for all the texts, while the human graders

were at least different in each collection. In an ideal condition, the model should represent and fit each individual grader in an isolated manner. However, the model is representing here a general good grader. Nevertheless, the error introduced by this generalization is not relevant for the purpose of the present work. It is also important to mention that only the aspects of quality related to text vocabulary, local coherence and grammatic and discourse structure have been considered in the study (since WM usage measures and model behavior depend on how the text is structured by sentences and their length). The extent to what the requested topics have been covered within the essays has not been taken into account.

Concluding remarks

The assessment of text quality is usually done by human graders. In order to avoid the subjectivity contained in their evaluations, the artificial scorers appeared as an objective alternative because they are based on quantitative linguistic features. The ultimate aim of using Natural Language is the communication among subjects. Consequently, if communication succeeds (it is correctly and easily comprehended), the quality of the original message will be high. Since the comprehension of natural language implies cognitive functions and resources, the text quality can be measured by characterizing the use of those cognitive resources and processes.

Given that current brain imaging techniques are not capable of acquiring a fine-grained measure of the underlying cognitive process, this paper has presented a computational cognitive model of Working Memory instead. The aim of this work is to unveil the effects of text quality and features in WM usage. The model emulates the cognitive dynamics of Working Memory during reading and allows monitoring memory usage and state. This way, the text quality can be correlated with some numerical factors that summarize and quantify the use of WM during the reading process of natural language texts. The results of the carried out experiments have shown how several factors concerning the average memory load and the average significance of concepts in memory determine the text score assigned by human scorers, thus proving the influence of text quality on the cognitive processing of the texts. More concretely, the results have revealed that as the score of the texts increases the activation of concepts in WM gets higher and more homogeneous.

Additionally, the effects of particular linguistic features of the text on WM have been studied. In this sense, the results have shown that the use of conditional, negative and connective expressions, and complex syntactic structures leads to an increment of the concepts and total activation in WM. In contrast, a high number of concrete words (non-generic) and highly qualified nouns produce a decrease in the number of concepts and total activation in WM. Besides, the incidence of negative and additive connectives together with the incidence of repetition of words are positively related to the average activation of each single concept in WM.

In the BICA context, the study and results presented here point to two broader conclusions. First, the presented model emulates a differentiated aspect of the mind, the so-called Bounded Cognition (Gigerenzer & Selten, 2002), which states that cognitive performance is influenced not only by the inherent cognitive capacity, but also by the information perceived and the environment. The inclusion of bounded cognition in BICA models means a step ahead in the achievement of the BICA Challenge. Second, the work described here means another example of how BICA models can feedback Psychology and Neuroscience, thus closing the interdisciplinary loop (Neuroscience/Psychology inform and inspired BICA models, which can be used to prove hypothesis from the former fields and generate new ones).

Other computational models of reading exist, but they search for an assessment of a theory of reading rather than for a biologically inspired approach. Most of them are based on connectionist networks inspired by the Construction–Integration model (Kintsch, 1988) and focus on different stages of reading and targets: in Rapaport and Shapiro (1999) a study of different ways of representing and understanding fiction in an associative net is presented. The interaction of different knowledge sources at sentence level during reading is treated in Mahesh, Eiselt, and Holbrook (1999). Representation of language for complex narrative understanding is studied in Domeshek, Jones, and Ram (1999). In Lange and Wharton (1993), the reminding process during reading is explained by inferencing and disambiguation and a connectionist model of episodic memory is proposed. A modification of the Construction–Integration model for narrative comprehension is presented in Langston, Trabasso, and Magliano (1999). In Meyer and Poon (2001) the importance of text structure and writing style for comprehension is highlighted. Even creativity is the target of studies in Moorman and Ram (1999) by the comprehension of novel concepts. The works just mentioned show that there is a high number of complex cognitive processes underlying reading. The model proposed here, CRIM, is a simple model that takes into account only a few those cognitive processes focused in WM. However, it is inspired by and slightly closer to humans than the other systems in the same application field.

For a better support of the results, it is planned to extend the collections of human-scored essays, and fit the model to each individual grader in an isolated manner, thus carrying out the same correlation analysis on the measures obtained from the models. Future work will also include a correlation study between the WM factors and the quantitative linguistic features of the texts modeling poor readers, instead of skilled graders, in order to unveil the effects of text quality under different cognitive capabilities. It is also planned to assess the plausibility of the model by modifying the low-quality texts according to the correlations found in this study in order to improve them in terms of the effect caused in WM. These modifications will be assessed afterwards by human graders, thus closing the loop human-model-improvement-human, which is a core aim of computational cognitive modeling.

Acknowledgment

This work has been funded by FGCSIC, Obra Social la Caixa and CSIC.

References

Awh, E., Vodel, E. K., & Oh, S. H. (2006). Interactions between attention and working memory. *Neuroscience, 139*, 201—208.

Beni, R. D., & Palladino, P. (2000). Intrusion errors in working memory tasks: Are they related to reading comprehension ability? *Learning and Individual Differences, 12*(2), 131—143.

Burstein, J., Leacock, C., & Swartz, R. (2001). Automated evaluation of essay and short answers. In M. Danson (Ed.), *Proceedings of the sixth international computer assisted assessment conference*. Loughborough, UK.

Calvo, M. G. (2005). Relative contribution of vocabulary knowledge and working memory span to elaborative inferences in reading. *Learning and Individual Differences, 15*(1), 53—65.

Cambria, E., Olsher, D., & Kwok, K., (2012). Sentic activation: A two-level affective common sense reasoning framework. In *Proceedings of AAAI conference on artificial intelligence*.

Cambria, E., Mazzocco, T., & Hussain, A. (2013). Application of multi-dimensional scaling and artificial neural networks for biologically inspired opinion mining. *Biologically Inspired Cognitive Architectures, 4*(0), 41—53.

Carrell, P. L. (1995). The effect of writers' personalities and raters' personalities on the holistic evaluation of writing. *Assessing Writing, 2*(2), 153—190.

Carretti, B., Borella, E., Cornoldi, C., & Beni, R. D. (2009). Role of working memory in explaining the performance of individuals with specific reading comprehension difficulties: A meta-analysis. *Learning and Individual Differences, 19*(2), 246—251.

Carretti, B., Cornoldi, C., Beni, R. D., & Romanò, M. (2005). Updating in working memory: A comparison of good and poor comprehenders. *Journal of Experimental Child Psychology, 91*(1), 45—66.

Christiansen, M. H., & Kirby, S. (2003). *Language evolution*. Oxford University Press.

Démonet, J.-F. (2005). The Dynamics of language-related brain images. *Neurocase, 11*(2), 148—150.

Domeshek, E., Jones, E., & Ram, A. (1999). Capturing the contents of complex narratives. In *Understanding language understanding: Computational models of reading* (pp. 73—105). MIT Press.

Duranti, A. (1997). *Linguistic anthropology*. Cambridge University Press.

Ericsson, K.A., & Kintsch, W. (1995). Long-term working memory. *Psychological review, 102*(2), 211—245.

Frase, L. T. (1969). Cybernetic control of memory while reading connected discourse. *Journal of Educational Psychology, 60*(1), 49—55.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality*. MIT Press.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2004). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweetm & C. E. Snow (Eds.), *Rethinking reading comprehension*. New York: Guilford Publications.

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

Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems, 15*(5), 22—37.

Higgins, D., & Burstein, J., (2006). In *Proceedings of the seventh international workshop on computational semantics (iwcs-7)*. Tilburg, The Netherlands.

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering, 12*(2), 145—159.

Jerrams-Smith, J., Soh, & V., Callear, D., (2001). Bridging gaps in computerized assessment of texts. In *Proceedings of the international conference on advanced learning technologies* (pp. 139—140).

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction—integration model. *Psychological Review, 95*(2), 163—182.

Lange, T. E. & Wharton, C. M. (1993). Dynamic memories: analysis of an integrated comprehension and episodic memory retrieval model. In: *Proceedings of international joint conference on artificial intelligence — IJCAI* (pp. 208—216).

Langston, M. C., Trabasso, T., & Magliano, J. (1999). A connectionist model of narrative comprehension. In *Understanding language understanding: Computational models of reading* (pp. 181—225). MIT Press.

Leacock, C. (2004). Statistical analysis of text in educational measurement. In C. F. G. Purnelle & A. Dister (Eds.), *Proceedings of the 7th international conference on textual data statistical analysis* (pp. 35—41). Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Lee, M. J., & Tedder, M. C. (2003). The effects of three different computer texts on readers' recall: Based on working memory capacity. *Computers in Human Behavior, 19*(6), 767—783.

Mahesh, K., Eiselt, K. P., & Holbrook, J. K. (1999). Sentence processing in understanding: Interaction and integration of knowledge sources. In *Understanding language understanding: Computational models of reading* (pp. 27—71). MIT Press.

Mason, O., & Grove-Stephenson, I. (2002). Automated free text marking with paperless school. In: Danson M., (Ed.), *Proceedings of the sixth international computer assisted assessment conference*. Loughborough, UK.

McDaniel, M. A., Hines, R. J., & Guynn, M. J. (2002). When text difficulty benefits less-skilled readers. *Journal of Memory and Language, 46*(3), 544—561.

Meyer, B. J. F., & Poon, L. W. (2001). Effects of structure strategy training and signaling on recall of text. *Journal of Educational Psychology, 93*, 141—159.

Moorman, K., & Ram, A. (1999). Creativity in reading: Understanding novel concepts. In A. Ram & K. Moorman (Eds.), *Understanding language understanding: Computational models of reading* (pp. 359—433). berlin: MIT Press.

Otten, M., & Berkum, J. V. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse? *Brain Research, 90*, 92—101.

Page, E. B. (1994). New computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*(2), 127—142.

Pinker, S. (1995). *The language instinct: How the mind creates language*. Perennial.

Pinker, S. (2000). *Words and rules: The ingredients of language*. Harper Perennial.

Rapaport, W. J., & Shapiro, S. C. (1999). Cognition and fiction: An introduction. In A. Ram & K. Moorman (Eds.), *Understanding language understanding: Computational models of reading* (pp. 11—25). MIT Press.

Saito, S., & Miyake, A. (2004). On the nature of forgetting and the processing—storage relationship in reading span performance. *Journal of Memory and Language, 50*(4), 425—443.

Serrano, J. I., del Castillo, M. D., & Iglesias, A. (2007). Characterizing individual interest by a computational model of reading. In R. Wang, F. Gu, & E. Shen (Eds.), *Advances in cognitive neurodynamics: Proceedings of international conference on cognitive neurodynamics ICCN 2007* (pp. 539—544). Shangahai, China: Springer.

Serrano, J. I., del Castillo, M. D., & Iglesias, A. (2009a). Assessing aspects of reading by a connectionist model. *Neurocomputing, 72*, 3659—3668.

Serrano, J. I., del Castillo, M. D., & Iglesias, A. (2009b). Dealing with written language semantics by a connectionist model of cognitive reading. *Neurocomputing, 72*, 713—725.

Serrano, J.I., del Castillo, M.D. & Iglesias, A. (2006). Cri, a computational model of cognitive reading for document indexing. In: I. I. Research & NAISO (Eds.), *Proceedings of brain-inspired cognitive systems (BICS06)* (pp. 66—72). Canada.

Sinclair, G. P., Healy, A., & Bourne, L. E. (1989). Facilitating text memory with additional processing opportunities in rapid sequential reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(3), 418—431.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education, 2*, 321—330.

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *The Journal of Technology, Learning, and Assessment, 6*(2), 2—28.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioural Research Methods, Instruments and Computers, 20*(1), 6—11.

Yussen, S. R., Stright, A. D., Glysh, R. L., Bonk, C. E., I-Chung, L., & Al-Sabaty, I. (1991). Learning and forgetting of narratives following good and poor text organization. *Contemporary Educational Psychology, 16*(4), 346—374.