

Computational cognitive models of summarization assessment skills

Benoît Lemaire, Sonia Mandin, Philippe Dessus, Guy Denhière

▶ **To cite this version:**

Benoît Lemaire, Sonia Mandin, Philippe Dessus, Guy Denhière. Computational cognitive models of summarization assessment skills. Proceedings of the 27th Annual Meeting of the Cognitive Science Society (CogSci' 2005), 2005, France. pp.1266-1271. hal-00372181

HAL Id: hal-00372181
https://hal.science/hal-00372181

Submitted on 31 Mar 2009

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

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

Computational Cognitive Models of Summarization Assessment Skills

Benoît Lemaire (Benoit.Lemaire@imag.fr)
Laboratoire Leibniz-IMAG (CNRS UMR 5522)
38031 Grenoble Cedex 1 FRANCE

Sonia Mandin (Sonia.Mandin@upmf-grenoble.fr)
L.S.E., University of Grenoble & IUFM
38040 Grenoble Cedex 9 FRANCE

Philippe Dessus (Philippe.Dessus@upmf-grenoble.fr)
L.S.E., University of Grenoble & IUFM
38040 Grenoble Cedex 9 FRANCE

Guy Denhière (denhiere@up.univ-mrs.fr)
L.P.C., University of Aix-Marseille & CNRS
13331 Marseille Cedex 1 FRANCE

Abstract

This paper presents a general computational cognitive model of the way a summary is assessed by teachers. It is based on models of two subprocesses: determining the importance of sentences and guessing the cognitive rules that the student may have used. All models are based on Latent Semantic Analysis, a computational model of the representation of the meaning of words and sentences. Models' performances are compared with data from an experiment conducted with 278 middle school students. The general model was implemented in a learning environment designed for helping students to write summaries.

Keywords: Summarization; Macrorules; Cognitive modeling; Computer environment; Latent Semantic Analysis.

Introduction

Summarizing information after reading a text is a very important and complex task. This ability can be both viewed as the cause of text comprehension (Thiede & Anderson, 2003) and its consequence (Brown & Day, 1983). This ability can be more precisely described by two distinct sub-models: – assessing the importance of the text read for selection purposes (Garner, 1987); – applying a set of macrorules like generalizing or deleting information (Brown & Day, 1983; Kintsch & van Dijk, 1978) that builds elements of the text macrostructure.

This paper is not concerned with the summarization process per se, whose result is a summary, but with the *summarization assessment process,* whose result is a *diagnosis* on a summary and possibly a global score. Actually, both processes probably share similar subprocesses. The teacher who is assessing a summary should have previously mentally assessed the importance of the text sentences. The difference probably lies in the fact that the teacher does not actually have to apply macrorules to construct the summary, but rather detects the use of these macrorules by the student.

The purpose of this paper is to design and test computational cognitive models of these two subprocesses. First, we will present various models of the way humans assess the importance of sentences in texts. Then, we will describe a method for automatically inferring the macrorules that humans could have used in producing a summary. Both models will be applied to a system that helps students to summarize texts.

Our process models are based on an underlying representation model, namely Latent Semantic Analysis (Landauer & Dumais, 1997), that can provide semantic relations between sentences or propositions.

The remainder of this paper is as follows. First, we will briefly present LSA, then four models of the sentence selection task, then a model of the way the use of macrorules could be detected in a summary, then a computer system that implements both text selection and macrorules detection.

A Representation of Sentences based on Latent Semantic Analysis

Modeling the activities involved in the summarization process should be done at the semantic level, especially for the purpose of cognitive modeling. We thus rely on LSA (Landauer, 2002), a powerful model for the representation of the meaning of words and sentences. LSA takes a huge corpus as input and yields a high-dimensional vector representation for each word. It is based on a singular value decomposition of a word x paragraph occurrence matrix, which implements the idea that words are given similar representations if they occur in similar contexts (not identical contexts, as LSA is often reduced to!). Such a vector representation is very convenient to give a representation to sentences that were not in the corpus: the meaning of a sentence is represented as a linear combination of its word vectors. Therefore, we can virtually take any sentence and give it a representation. The second advantage of the vector representation is that it is straightforward to compute the similarity between vectors, usually by means of the cosine function. Our models largely rely on this LSA measure of semantic similarity.

The corpus from which the semantic space is built plays a large role, especially for the purpose of cognitive modeling. If the semantic similarity between words or pieces of text is meant to model human associations in semantic memory, then the corpus should correspond as closely as possible to the kind and amount of text humans are exposed to.

The LSA semantic space which was used in all four models was built from a 13 million word corpus composed of three sub-corpora:

- a 3.3 million word corpus representing the kind of texts participants were exposed to during their childhood (Denhière & Lemaire, 2004);
- a 5 million word corpus composed of novels;
- a 5 million word corpus composed of newspaper articles.

This huge corpus was processed by LSA and all words were represented as vectors in a 300-dimensional space. We will now successively present the two subprocesses that are part of the summary assessment skills.

Subprocess 1: Sentence Selection

This first set of models seeks to describe the way humans assess the importance of sentences in a text. These models are not specific to the summarization assessment process, they are probably the same as the summarization process itself.

Our four models manage differently the cognitive processes which may be involved in identifying the most important sentences in a text. The first one postulates that, in so doing, we compare each sentence to the entire text (E. Kintsch et al., 2000). The second model considers that the reader would consider as being important those sentences that are highly connected to the others. We borrowed this idea from Kintsch's (2002) notion of sentence typicality, which is the semantic relation between a sentence and all other sentences in its text section. The third model postulates that the reader is rather aware of coherence gaps between two blocks of sentences (Foltz, Kintsch & Landauer, 1998). The last model views the main idea selection as the result of the sentence by sentence comprehension of the text, by the way of the Construction-Integration model (Kintsch, 1998).

These four computational models will be successively presented and all simulations will be compared to human data. We first present the experiment which provided these data.

Human Experiment

We carried out a human experiment to collect empirical data to which each of our computational models of sentence selection was to be compared. We chose participants above grade 7 for their adult-like capability to rate important passages (Hidi & Anderson, 1986). A total of 278 middle school students (grades 8 to 11, see Table 1 for the distribution) were given a single-page text among two: an expository text, entitled "Elephants' drugstore" or a narrative text, entitled "Miguel". These texts respectively contained 523 and 382 words (18 and 24 sentences). The average number of words per sentence was 29.06 (SD = 14.66) in the first text, and 15.92 (SD = 8.22) in the second one. The expository text was selected because of the lack of participants' prior knowledge in this domain. The task was to read the text and to "underline three to five sentences on the sheet, that seemed to be the most important". The underlined sentences were then compared to the set of sentences selected by the four following models. Any sentence partially underlined by participants was categorized as entirely underlined. Our tables will indicate results by grade because we were looking for possible differences. Since we do not know whether the differences found are due to a school or class effect, they are hardly interpretable. We therefore discuss only on the overall results.

Table 1: Participants' distribution between grades and text read.

Text/Grade	8th	9th	10th	11th	Total
Narrative	25	39	55	19	138
Expository	25	39	54	22	140

Model 1: Important Sentences have a High Semantic Similarity with Text

In our first model, we postulate that important sentences have a high semantic similarity with the whole text. The two texts and all their sentences were represented as vectors in the semantic space mentioned earlier. All sentences were assigned a measure of importance which was their cosine with the text. Correlations with human data are presented in Table 2. Results show a good adequacy for the expository text, much better than for the narrative text. It is worth noting that such a simple model of importance assessment could so well mimic human judgments.

Table 2: Within-grade correlations between model 1 and human data.

Text/Grade	8th	9th	10th	11th	Overall
Narrative (N = 24)	.37	.18	.34	.28	.31
Expository (N = 18)	.52*	.70**	.59**	.58*	.64**

*p < .05; **p < .01

One reason why the narrative text does not yield good results could be due to the LSA rule of compositionality: the meaning of the whole text is a linear combination of the meaning of its sentences. A text is therefore an aggregate structure which tends to flatten the individual meanings of its sentences. It is not a problem with expository texts for which all sentences are related to the general theme. For instance, all sentences of our expository text have to do with elephants. Narrative texts are quite different. They usually have a "plot" (Pinto Molina, 1995). Sentences must be linked to that plot, but not necessarily to an "average meaning" of the story. Therefore, this model might not be adequate for narrative texts.

Model 2: Important Sentences are Highly Connected to Other Sentences

The second model is more fine-grained. Instead of considering the text as a whole, it breaks it into sentences. It is based on the idea that important sentences are highly connected to others. The degree of connectivity between two sentences is defined as the cosine of their vectors. Therefore, we define the importance of a sentence as the number of other sentences whose cosine with the current one is above a given threshold (.12 in this study). Table 3 displays the correlations with human data. The expository text was still better than the narrative one and overall results are better than for model 1.

Table 3: Within-grade correlations between model 2 and human data.

Text/Grade	8th	9th	10th	11th	Overall
Narrative (N = 24)	.34	.10	.48*	.37	.37
Expository (N = 18)	.53*	.68**	.62**	.65**	.66**

*p < .05; **p < .01

Model 3: Important Sentences Belong to Coherent Blocks

Another dimension of the structure of a text to be summarized is its coherence. This notion expresses the amount of relatedness between text units (e.g., sentences, paragraphs), provided that the more coherent are parts of a text, the more they could be selected as the most important ideas from this text (i.e., the macrostructure of the text). This dimension is related to the capability to recognize connections between adjacent elements of the text.

The third model we tested used coherence between sentences to predict their importance. This model postulates that, given a block of sentences, important ones are merely placed at the beginning and at the end of the block (Baxendale, 1958; Williams et al., 1984). We defined a coherent block as a consecutive chain of sentences separated from others by a coherence gap.

Barzilay and Elhadad (1997) used such a procedure for automatically extracting the most important segments of a text. First, they constructed a net of semantically connected words, and then they carried out three heuristics. The third is especially important for our purpose: "[the central topic is] a cluster of successive segments with high density of chain members". This procedure resembles ours because we designed a model able to capture successive segments of texts that also are highly coherent.

The procedure used in the model 3 is as follows. This model is a generalization of the result presented above: the first and last sentences of a paragraph are often considered important. The notion of paragraph can be extended to any block of sentences (e.g., text) and our model becomes: first and last sentences of a coherent block of text are considered important. The two source texts were thus processed in order to find out coherent blocks of sentences, that is, any sequence of inter-sentence similarities above an arbitrary threshold (.11 in this study). We used LSA to compute these inter-sentence similarities following Foltz et al. (1998) method: similarity between two adjoining sentences is the cosine of their vectors. The first and last sentences of each text were obviously also selected.

For example, given the sentences 12, 13, 14 with inter-sentence similarities respectively of .1 (between 12 and 13), .5 (between 13 and 14) and .1 (between 14 and 15), sentences 13 and 14 are considered part of a coherent block. In order to match results from the model with human ones, selected sentences were coded 1, others were coded 0.

Table 4 shows that the comparison between human selection and data from this model yields significant values for the expository text, but rather poor ones for the narrative text, comparable to the data from the first model.

Table 4: Within-grade correlations between model 3 and human data.

Text/Grade	8th	9th	10th	11th	Overall
Narrative (N = 24)	.22	.23	.35	.27	.30
Expository (N = 18)	.48*	.57*	.41	.54*	.51*

*p < .05

Comparing the three previous models, we could say that model 2 is the best: it is the most fine-grained and it takes into account all text sentences instead of just adjacent ones. It could be a reason why this kind of model is widely used in the field of automatic summarization.

Model 4: Important Sentences have High Activation Values in a Simulation of their Comprehension

The fourth model attempts to integrate information among sentences. Instead of considering each sentence in isolation and compare it to the whole text (model 1), other sentences (model 2), or the preceding and following sentences (model 3), it takes into account the time course of sentence comprehension. The importance of sentences is given by their activation value at the end of the whole text processing. This model is based on the Construction-Integration model of text comprehension (Kintsch, 1998). It relies on three different memory structures:

- a working memory which is a set of concepts (words) or propositions as well as their activation values. Items come from the text itself, the semantic memory or the episodic memory.
- a semantic memory, simulated by LSA, which can provide associates to concepts or propositions.
- an episodic memory, which keeps track of all concepts or propositions that occur in working memory, as well as their activation value which tend to decay over time.

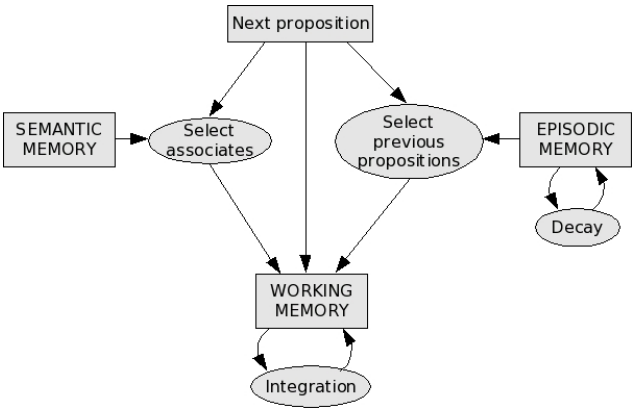


Figure 1: Architecture of model 4.

Figure 1 describes the architecture. The text is processed proposition by proposition (sentence by sentence in this study for the sake of homogeneity). Each proposition activates semantic neighbors from semantic memory (for instance, *the bee is sucking nectar from a flower* activates *honey* and *hive*). They are all added to working memory. Concepts or propositions from episodic memory can also be added to working memory if they are close enough (and activated enough) to one of the current elements. The integration algorithm defined by Kintsch (1998) is then applied to provide activation values to elements and to rule out irrelevant ones. It is based on a matrix of semantic similarities also provided by LSA. For instance, a concept like *florist* that could have been activated from semantic memory would be given a low activation value and removed from working memory in the previous context of the *bees*. However, a proposition that could have occurred previously in the text, like *How the honey is made w*ould be retrieved from episodic memory, given a high activation value and kept in working memory.

This model was implemented in Perl and hooked up to LSA. It is available on demand for academic purpose.

Our two texts were processed with this model and the final activation values of all text sentences were collected. These values were then compared with the participants' judgments of importance (see Table 5). It turns out that correlations are more homogeneous. They are even better than previously for the narrative text and worse for the expositive one.

Table 5: Within-grade correlations between model 4 and human data.

Text/Grade	8th	9th	10th	11th	Overall
Narrative (N = 24)	.42*	.42*	.39	.41*	.43*
Expository (N = 18)	.30	.31	.39	.37	.37

*p < .05

Our four models follow a progression in the way the sentences are assessed. In model 1, sentences are compared to the whole text. Model 2 is intended to be more fine-grained since sentences are compared to other sentences, no matter their location in the text. Model 3 operates at the same level, but considers only adjacent sentences (the previous one and the next one). Model 4 is another refinement since it is entirely dependent on the sentence order by automatically extracting the text macrostructure. This last model is quite different from the three others. It is based on a representation of different memory structures and could even retrieve concepts that were not in the text. These features are especially necessary for processing narrative texts since the connections between sentences are less obvious than in expository texts. All the sentences of our expository text deal with the same topic, whereas the domain of the narrative text is much broader. This could explain why model 4 produced better results with the narrative text. For the purpose of assessment, we compared our models to a simpler model, namely the *Microsoft Word* summary generator. Its overall correlations with human data appear to be lower than previous ones (r = .45 for the narrative text; and r = –.30 for the expository one).

Subprocess 2: Detecting the Use of Macrorules

We are now concerned with the second subprocess of the summarization assessment process. Besides identifying important sentences in the text, the teacher is engaged in the detection of the student's strategy, which is viewed as the application of adequate macrorules.

Macrorules are the core of the cognitive processes involved in the summarization activity (Kintsch & van Dijk, 1978).These authors described three macrorules:

- *deletion*, where each proposition (in our case, sentence) that either contains minor, redundant or unrelated details may be deleted;
- *generalization*, where "each sequence of propositions may be substituted by the general proposition denoting an immediate superset" (id., p. 366);
- *construction*, where "each sequence of propositions may be substituted by a proposition denoting a global fact of which the facts denoted by the microstructure propositions are normal conditions, components, or consequences". (ibid.)

We did not implement these macrorules, which is quite a hard task (Kintsch, 2002), but rather modeled the detection of their usage by the student. Given a text and its summary, a teacher is able to infer that this particular sentence in the text has been deleted, that this summary sentence is a generalization, and so on. Our goal is to account for that task.

We also designed three additional macrorules which more extensively describe the operations on the source text (after Brown & Day, 1983):

- *paraphrase,* which consists in writing down a semantically similar sentence;
- *copy*, for which the resulting sentence is copied almost exactly;
- *off-the-subject*, when a sentence is added without being related to the subject.

As for the selection subprocess, that subprocess was operationalized through the use of LSA: each sentence of the summary is semantically compared with each sentence of the source text (ST). For instance, a sentence of the ST would be considered as *deleted* if no sentence of the summary is sufficiently close to it. In the same way, a *generalized* sentence is a sentence of the summary that is sufficiently close to more than one sentence of the ST.

It is then necessary to operationalize this very notion of closeness: how close to another is a "rather close" sentence? Three similarity thresholds have been used corresponding to the following similarities: not enough similarity (cosine < .2), low similarity (.2 < cosine < .5), good similarity (.5 < cosine < .8), too high similarity (cosine > .8). The comparison of each sentence from the summary with all N text sentences results in a distribution of N similarities among these four categories. It is that distribution that permits the detection of the student strategy.

Figure 2 shows an example of a distribution of similarities with a given summary sentence. Five text sentences are semantically too far from the summary sentence (# 13, 12, 6, 7 and 4), eight share some relation with it (# 10, 3, 5, 8, 14, 16, 1 and 15), three have good similarity with it (# 11, 9 and 2) and none is almost identical. This distribution indicates that the summary sentence is probably a generalization, since there are three sentences of the text that are highly similar to it.

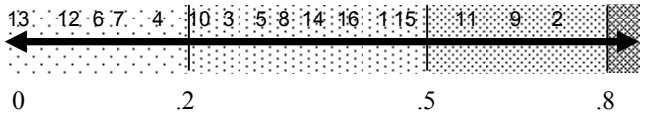


Figure 2: Representation of the comparisons between a given summary sentence and each source text sentence (represented by numbers).

More precisely, the categorization is the following. Let $Q_i = (x_1, x_2, x_3, x_4)$ be the distribution of similarities over the four categories. In the previous example, $Q_i = (5,8,3,0)$. The number of sentences of the text is $x_1+x_2+x_3+x_4$. If we consider that '?' indicates an unspecified value, we will say that a sentence R_i of the summary is:

- a *copy* if $Q_i = (?, ?, ?, N)$, $N \geq 1$ (there is at least one sentence of ST which is very close to R_i);
- a *generalization* if $Q_i = (?, ?, N, 0)$, $N \geq 2$ (there are several sentences of the ST that are close to R_i);
- a *paraphrase* if $Q_i = (?, ?, 1, 0)$ (there is only one sentence of the TS that is close to R_i);
- a *construction* if $Q_i = (?, N, 0, 0)$, $N \geq 1$ (no sentences of TS are close to R_i, but at least one of them is somehow related to R_i);
- *off-the-subject* if $Q_i = (?, 0, 0, 0)$ (all sentences of TS are unrelated to R_i).

Obviously, the length of the summary sentence also plays an important role in the diagnosis: each summarized sentence has to be shorter than the reference sentence. This model is currently a theoretical model. It has been implemented in a computer program but has not been tested yet. The reason is that its validation is much more difficult than for the previous model. One way to do that would be to compare the results of the model to teachers' categorizations of summary sentences.

Application: a Learning Environment to Help Students to Summarize Texts

Since our models are computational models, they have been integrated into a learning environment. The goal is to rely on the information we identified (importance of sentences, cognitive macrorules inferred) to provide students with several prompts that can help them to produce their summaries. Our system resembles other tutoring systems (Halpin et al., 2004; Wade-Stein & Kintsch, 2004) in which texts rewritten by students are automatically compared to the original text. Prompts are then given to the child according to the quality of the matching.

Macrorules are more or less mentally elaborated and difficult to perform depending on the age and the competence of the student; their application also has an effect on the production of the summary (Brown & Day, 1983). Furthermore, research showed that training students to identify main ideas as well as to apply higher-order macrorules (e.g., generalization, construction) led them to increase the quality of written summaries (Casazza, 1993).
Once the macrorules have been detected, it is useful to prompt the student to use the most elaborated ones. An overall score can also be delivered taking into account this application. For instance, if a student was generalizing a ST sentence using too many words, the student would be warned: "This generalized sentence might be longer than the corresponding ST sentences". It is worth noting that the diagnosis also takes into account the importance of the ranking of the sentences. For instance, if the student has paraphrased an unimportant sentence, she or he would be warned.

Our environment implements the two subprocesses described previously. Assessing the importance of sentences in the source text is performed by an implementation of model 1. Detecting the use of macrorules and providing a diagnosis was implemented from the theoretical model we presented previously. The interaction with the student is as follows. First, the student is provided with a source text. After reading it, the student writes out a summary of the text in another window. Secondly, at any time the student can get an assessment of the summary. This feedback may either highlight sentences depending on whether they are adequate or not, or deliver diagnostic messages about the macrorules the student applied. It is worth noting that our environment does not generate any "expert" summary to be compared to the students' summary. It rather diagnoses whether the student actually applied the macrorules to the different sentences of the summary.

Conclusion

The main asset of the learning environment we have just been presenting is its cognitive foundations. The summarizing process engages numerous complex cognitive skills that have to be assessed in order to assess a summary.

We considered a two-step process: assessing the importance of sentences and detecting the student use of macrorules. Although the first subprocess has been subject to empirical validation, the second remains to be confronted to human data. We thus plan to ask teachers to detect and diagnose macrorule application after reading student's summaries. This comparison against human data could help tackle a major issue: LSA-based models often rely on similarity thresholds to decide between two alternatives (sentences are coherent or not, words are semantically related or not, etc.). However, it is quite hard to set the value of those thresholds. They are often arbitrarily determined and we plan instead to perform a fine tuning by a comparison to human data.

The cognitive skills involved in the summarization process probably depends on the nature of the text read. Our follow-up investigations will take into account this major difference between narrative and expository texts.

Acknowledgements

This research was supported by a grant from the French Research Ministry under a *Cognitique* project leaded by Guy Denhière. We would like to thank Beulah Henry and Françoise Raby for their thoughtful comments on an earlier version of this paper; and Céline Marchais for her help during the experiments. We also thank teachers who kindly accepted to pass the text selection experiment in their classes.

References

Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. *Proc. ISTS'97,* Madrid.

Baxendale, P. B. (1958). Machine-made index for technical literature – An experiment. *IBM J. Res. Dev.*, *2*, 354-365.

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *J. Verb. Learn. Verb. Behav., 22*, 1-14.

Casazza, M. E. (1993). Using a model of direct instruction to teach summary writing in a college reading class. *Journal of Reading, 37*, 202-208.

Denhière, G., & Lemaire, B. (2004). A Computational Model of Children's Semantic Memory, in *Proc. 26th Annual Meeting of the Cognitive Science Society (CogSci'2004)*, 297-302.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25*, 285-307.

Garner, R. (1987). *Metacognition and Reading Comprehension*. Norwood: Ablex.

Halpin, H., Moore, J. D., & Robertson, J. (2004). Automatic analysis of plot for story rewriting. *Proc. EMNLP 2004*. Barcelona, Spain.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Tasks demands, cognitive operations, and implications for instruction. *Rev. Educ. Res., 56*, 473-493.

Kintsch, W. (1998). *Comprehension, a Paradigm for Cognition*. Cambridge: Cambridge University Press.

Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.), *Thematics: Interdisciplinary Studies* (pp. 157-170). Amsterdam: Benjamins.

Kintsch, E., Steinhart, D., Stahl, G., LSA Research Group, Matthews, C., & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learn. Env., 8*, 87-109.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychol. Rev., 85*, 363-394.

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *Psychol. Learn. Motivation, 41*, 43-84.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychol. Rev., 104*, 211-240.

Pinto Molina, M. (1995). Documentary abstracting : toward a methodological model. *J. Am. Soc. Inform. Sci., 46*, 225-234.

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemp. Educ. Psychol., 28*, 129-160.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*, 333-362.

Williams, J. P., Taylor, M. B., & de Cani, J. S. (1984). Constructing macrostructure for expository text. *J. Educ. Psychol.*, *76*, 1065-1075.