DEPARTMENT OF MANAGEMENT

The effect of errors in the text produced so far
Strategy decisions based on error span, input mode, and lexicality

Mariëlle Leijten, Isabelle De Ridder, Sarah Ransdell & Luuk Van Waes

UNIVERSITY OF ANTWERP
Faculty of Applied Economics

Stadscampus
Prinsstraat 13, B.213
BE-2000 Antwerpen
Tel. +32 (0)3 220 40 32
Fax +32 (0)3 220 47 99
http://www.ua.ac.be/tew

FACULTY OF APPLIED ECONOMICS

DEPARTMENT OF MANAGEMENT

The effect of errors in the text produced so far Strategy decisions based on error span, input mode, and lexicality

Mariëlle Leijten, Isabelle De Ridder, Sarah Ransdell & Luuk Van Waes

RESEARCH PAPER 2007-009
MAY 2007

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.213
phone: (32) 3 220 40 32
fax: (32) 3 220 47 99
e-mail: joeri.nys@ua.ac.be

The papers can be also found at our website:
www.ua.ac.be/tew
(research > working papers)

D/2007/1169/009

The effect of errors in the text produced so far

Strategy decisions based on error span, input mode, and lexicality

Mariëlle Leijten*, Isabelle De Ridder°, Sarah Ransdell• & Luuk Van Waes*

* University of Antwerp
° The Flemish Education Council
• Nova Southeastern University

May 21, 2007

Contact: Mariëlle Leijten, University of Antwerp, Faculty of Applied Economics, Department of Management, Prinsstraat 13, BE-2000 Antwerp, Belgium (marielle.leijten@ua.ac.be)

Abstract: Error analysis involves detecting, diagnosing and correcting discrepancies between the text produced so far (TPSF) and the writers mental representation of what the text should be. While many factors determine the choice of strategy, cognitive effort is a major contributor to this choice. This research shows how cognitive effort during error analysis affects strategy choice and success as measured by a series of online text production measures. Text production is shown to be influenced most by error span, i.e. whether or not the error spans more or less than two characters. Next, it is influenced by input mode, that is whether or not the error has been generated by speech recognition or keyboard, and finally by lexicality, i.e. whether or not the error comprises an existing word. Correction of larger error spans are corrected more successful smaller errors. Writers impose a wise speed accuracy tradeoff during large error spans since correction is better, but preparation times and production times take longer, and interference reaction times are slower. During large error spans, there is a tendency to opt for error correction first, especially when errors occurred in the condition in which the TPSF is not preceded by speech. In general the addition of speech frees the cognitive demands of writing: shorter preparation and reaction times. Writers also opt more often to continue text production when the TPSF is presented auditory first.

Keywords: Cognitive effort, dictation, dual task technique, error analysis, speech recognition, text produced so far (TPSF), text production, technology of writing, working memory.

The effect of errors in the text produced so far

Strategy decisions based on error span, input mode, and lexicality

Abstract: Error analysis involves detecting, diagnosing and correcting discrepancies between the text produced so far (TPSF) and the writers mental representation of what the text should be. While many factors determine the choice of strategy, cognitive effort is a major contributor to this choice. This research shows how cognitive effort during error analysis affects strategy choice and success as measured by a series of online text production measures. Text production is shown to be influenced most by error span, i.e. whether or not the error spans more or less than two characters. Next, it is influenced by input mode, that is whether or not the error has been generated by speech recognition or keyboard, and finally by lexicality, i.e. whether or not the error comprises an existing word. Correction of larger error spans are corrected more successful smaller errors. Writers impose a wise speed accuracy tradeoff during large error spans since correction is better, but preparation times and production times take longer, and interference reaction times are slower. During large error spans, there is a tendency to opt for error correction first, especially when errors occurred in the condition in which the TPSF is not preceded by speech. In general the addition of speech frees the cognitive demands of writing: shorter preparation and reaction times. Writers also opt more often to continue text production when the TPSF is presented auditory first.

Keywords: Cognitive effort, dictation, dual task technique, error analysis, speech recognition, text produced so far (TPSF), text production, technology of writing, working memory.

1 Introduction

To write well is to write with an eye for change. Modern writing technology allows the writer to produce and change the text easily via keyboard based word processing. As a consequence, a high degree of non-linearity is characteristic for these writing processes. In the late 90’s speech recognition emerged as a writing medium. This medium is a hybrid writing mode that combines characteristics of classical dictating and word processing. The main strength of speech recognition lies in the combination of high speed text composition and the appearance of the text produced so far (TPSF) on the screen. However, writing with speech recognition does not yet result in a 100% faultless text on screen. For instance, when a writer dictates ‘various’ it can be recognized as ‘vary us’. This kind of (semantic) errors require extra monitoring and make it more difficult to benefit from the speed of composition (Honeycutt, 2003). Consequently, writers that use speech recognition for text production must revise intensively. Even more than other writers they need to ‘write with an eye for change’.

Revision during writing involves error analysis, comprising error detection, diagnosing and correction. This process has received much attention in cognitive science (cf. Rabbitt, 1978; Rabbitt, Cummings, & Vyas, 1978; Sternberg, 1969) and, more recently in computer based writing research (Hacker, 1997; Hacker, Plumb, Butterfield, Quathamer et al., 1994; Larigauderie, Gaonac’h, & Lacroix, 1998; Piolat, Roussey, Olive, & Amada, 2004). Even more recently, Leijten and Van Waes (2005b) reported on various error correction strategies of professional writers that were novice speech recognition users. The speech recognition users seemed to switch frequently from detection to correction, rather than continuing to write, resulting in a quite non-linear writing process. However, this observation did not hold for all the writers. A case study showed that one writer preferred to correct errors in the TPSF *immediately* and that the other writer showed a preference to *delay* error correction, with the exception of typical keyboard errors.

Writers always need to make strategy decisions to deal with the errors in the TPSF. In the speech recognition mode, however, another process emerges. The writer constantly has to answer the question ‘Is the text produced so far correctly presented?’ In other words, is the text that the writer has dictated to the speech recognizer correctly presented in the text that appears on the screen (see Figure 1)?

The text dictated via speech recognition can either be presented correctly or not. If it is correct, the writer can continue with text production or plan on revising the text at a later point. Or the text can be presented incorrectly because a technical misrecognition occurred. In that instance, the writer may or may not detect the error in the TPSF. If the writer does, he can choose again between text production and reviewing. If a technical problem is detected by the writer, he can then ignore the problem, solve it at a later stage, or correct the error immediately. In the latter case, he can either perform the technical revision immediately or he can train the speech recognizer for future reference. The speech recognizer does not necessarily ‘know’ every word that writers use and therefore one can make the software smarter by adding words to the dictionary and by training the pronunciation of words. Ultimately, speech recognition can lead to a strategy of delaying (minimal) error correction.

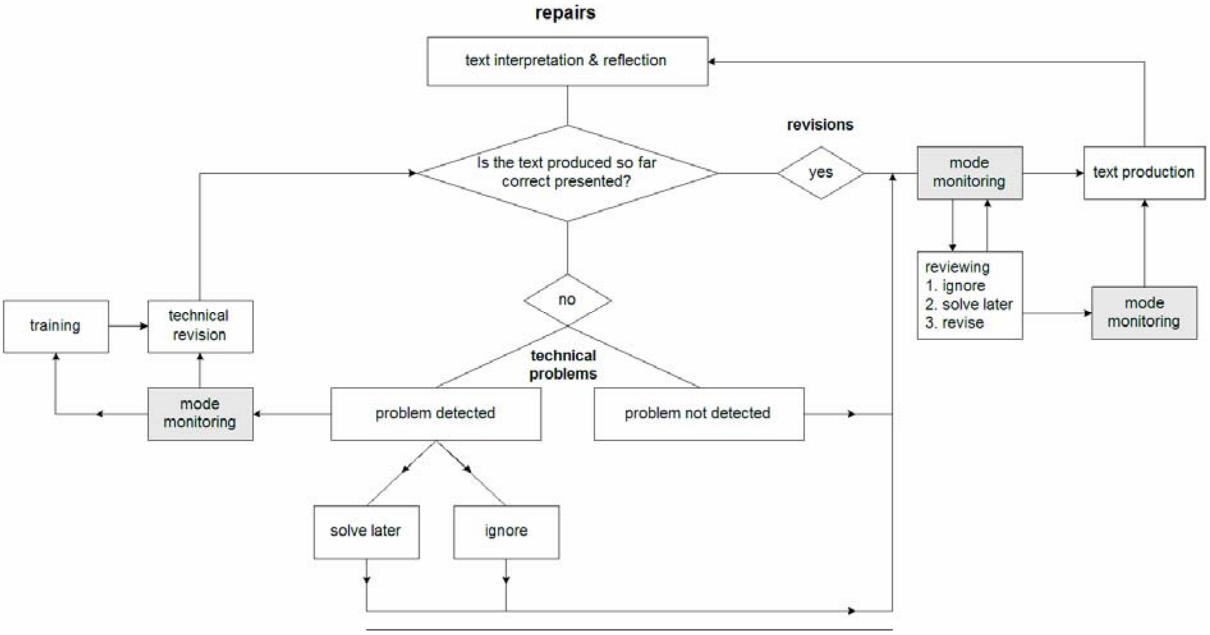


Figure 1. Speech recognition and the text produced so far.

A related quantitative study (Leijten & Van Waes, 2006) showed that writers not only differed in the way they repair errors but also in the number and type of errors they solve immediately. Some participants solved almost all the keyboard & mouse errors in the text immediately – possibly because there was no need to switch writing modes to solve these repairs – while they were much more tolerant of speech recognition errors. Other writers, however, are less tolerant of this type of error and often immediately solve almost all larger errors, typical speech recognition errors, and errors that were located at the point of utterance. However, all writers did solve errors involving nonexistent words immediately. They seemed intolerant of these errors in their texts.

Is it strange that a writer, who prefers a first time final draft, at the same time delays to correct a few errors? Are those errors not solved on purpose or are they overlooked? A possible answer could be that smaller errors and errors in the beginning of the sentence are easier to miss. Earlier research has already shown that rereading of the TPSF with the intention to further generate and formulate text is characterized by a high degree of success (Blau, 1983). Writers in those circumstances do not really evaluate the correctness of their text, but only observe the 'Gestalt' of what has been written as a trigger to further text production. So, on the one hand the interaction with the text on the screen can lead to a highly recursive writing process in which every error is repaired almost immediately, but on the other hand it can also lead to a less recursive writing process in which errors are corrected at the end of a paragraph or a text and left unnoticed at the point of utterance.

The objective of the present study is to explain differences in revising behavior: why are certain errors immediately corrected, while others are delayed? We assume that working memory plays an important role in this decision process. Therefore, we describe the differences in cognitive load caused by various error types. In other words, some error types request too much attention of a writer, and need to be corrected immediately, before text production can be continued. Other errors can remain in the memory of the writer and do not need to be solved immediately. It is plausible that error correction strategies of writers are affected by working memory.

In the following sections we describe the requirements for successful error detection, the effect of the TPSF on the writing process and finally we elaborate on the influence of working memory in writing processes.

1.1 Successful error detection

Error detection is vastly improved if one can anticipate the requirements for fixing errors (Rabbitt, Cummings, & Vyas, 1978). During writing, errors can be any discrepancy between the TPSF and one's mental representation of how the text should be. Errors come in a wide variety of types and some are easier to process than others.

Larigauderie, Gaonac'h and Lacroix (1998) found that central executive processes in working memory (cf. section 4.3) are involved in detecting semantic and syntactical errors, but less so for typographical errors.

Furthermore, they found that the disruption of the phonological loop mainly affected processing above the word level. They also found that greater processing spans, ranging from one word, several words within a clause, to words across clause boundaries respectively, required more memory resources than smaller spans. These two variables, error type and processing span were additive in their effects on successful error correction. In the writing task Larigauderie et al. (1998) used, a page long text was presented including errors of many types not isolated by an experimental design. In the present experimental study, we present error types that naturally occur in a typical writing task to determine strategy decisions writers make at the point of utterance when hearing and/or seeing text.

Hacker et al. (1994) found that writers first need to know how to correct a wide range of errors (meaning-based, grammar-based, or spelling-based errors) to detect them accurately. However, if an error is a simple typo, it is easier to detect than a meaning-based error because the latter requires text comprehension. Not only were spelling errors better detected, their detection also predicted correction. Not surprisingly, writing time, error type determination, along with the writer's linguistic knowledge, and knowledge of the text topic, facilitate error detection and correction.

At present, the jury is still out on how writing technologies such as speech recognition technology impact successful error detection. Detailed online records of writing by keyboard and speech recognition software will reveal the most common types of errors created by both, and the ways in which writers change their strategies to accommodate them.

1.2 The effect of the text produced so far on the writing process

The TPSF may play a different role in writing with speech recognition than it does computer-based word processing. Since speech recognition combines characteristics of both classical dictating and computer writing, it is useful to describe speech recognition as a hybrid writing mode (see Figure 2). That is why we position speech recognition between classical dictating and computer based word-processing. Similar to classical dictating speech recognition, texts are also audible via text-to-speech. Like in computer writing the TPSF in the speech recognition mode is visible on the screen (1). As in computer writing, the emerging text appears almost immediately on the screen, not letter by letter, but in text segments or phrases (2). After training, the text on screen is a more or less correct representation of what has been dictated (3).

We consider the presentation of the classical dictating device to be always correct because the audio on the tape is identical to the input; of course, the transition to paper is dependent on the typist[1]. In computer writing, the on screen representation as classified as (semi) correct because of the typing errors that may occur. However, the misrecognitions of the speech recognizer - and consequently the incorrect representations on the screen – are of a different kind. There is a possible overt conflict between the TPSF and speech in speech recognition and a possible covert conflict in computer writing (4).

Classical dictating	Speech recognition	Computer writing
1. invisible text (audible after rewind) 2. no simultaneous feedback 3. correct presentation 4. no conflict in presentation	1. visible and audible text 2. (semi) simultaneous feedback 3. (semi) correct presentation 4. overt conflict between TPSF and speech	1. visible text on screen 2. simultaneous visual feedback 3. (semi) correct presentation 4. covert conflict between TPSF and text as intended

Figure 2. Hybrid characteristics of speech recognition.

These characteristics lead to differences in the writing process between classical dictating and computer writing. Previous studies show that classical dictating is characterized by a high degree of linearity in the text production (Schilperoord, 1996). Writers dictate sentences or phrases one after the other and only few revisions are made. The only revising usually taking place is a mental revision before the text is dictated to the recorder. The computer writing process is typically characterized by a high degree of non-linearity (Severinson Eklundh, 1994; Van Waes & Schellens, 2003). Most computer writers consider the paragraph, or even a sentence, as a unit that is planned, formulated, reviewed and revised in short recursive episodes (Van den Bergh & Rijlaarsdam, 1996). The constant

[1] For a more detailed description of speech recognition and text-to-speech we would like to refer to Honeycutt (2003), MacArthur (2006) and Quinlan (2006).

feedback on the screen offers them the possibility to revise extensively, without losing the overview of the final text (Haas, 1989a, 1989b; Honeycutt, 2003).

So, in contrast to the traditional dictating mode, writers using speech technology receive immediate written feedback on the computer screen that may overtly conflict with the dictated TPSF. This previously mentioned technical characteristic creates the possibility to review the text in all stages of the writing process either by speech or by the complementary use of keyboard (without speech), inviting non-linearity (4).

1.3 Working memory and writing

Revising in general and error correction in particular can be seen as cognitive demanding activities. Almost every article on writing includes a paraphrase on 'writing is cognitively demanding'. These high demanding activities have been described in several models on working memory (Baddeley, 1986; Baddeley & Hitch, 1974; Ericsson & Kintsch, 1995; Kellogg, 1996; McCutchen, 1996; Shah & Miyake, 1996). Writers need to juggle the constraints of the several subprocesses. As Torrance and Galbraith (2006, p. 12) state:

> Finally, we have suggested that although some aspects of the writing process can be strategically controlled, others, such as the need to suppress irrelevant information or the need to re-read to refresh transient memory, arise as a consequence of a cycle of processing as it occurs on-line. No matter how skilled we are at managing the writing process, there is an irreducible core of potential conflicts. Writing will always be a struggle to reconcile competing demands. Writers have – motivationally – to accept this if they are to get the task done. (p. 12)

Kellogg integrated the six basic writing subprocesses with the working memory model of Baddeley (Kellogg, 1996, 2004; Levy & Marek, 1999). This model on working memory was developed after a range of experiments on working memory processing. Via the dual task paradigm (cf. method section 2.4) Baddeley found that the working memory consist of several subsystems. This main finding led to the development of a tripartite model (Baddeley, 1986). In short, the model consists of a central executive and two slave subsystems, the visuo-spatial sketchpad and the phonological loop. The visuo-spatial sketchpad stores visual and spatial information (i.c. visual TPSF) and the phonological loop stores verbal and auditory information (i.c. auditory TPSF). The central executive manages both parts.

The writing subprocesses that are most relevant to this study are reading and editing. Kellogg states that reading is related to the central executive and the phonological loop. Editing is related to the central executive. Figure 3 shows all the relations. Of course, editing can also happen prior to text execution. In this experiment we focus only on editing of the TPSF (cf. difference between internal and external revisions, Lindgren & Sullivan, 2006). Kellogg states that editing signals errors in the output of planning, translating, programming, and executing. Then, feedback about the error to the appropriate process is needed. A recursive pattern of the above-mentioned planning to executing process is put into action. This recursive process can occur immediately after production of the error, but may also be delayed. The strategy adopted by the writer for allocating working memory to monitoring versus formulation and execution affects the decision process of correcting immediately or delaying error correction (Kellogg, 2004).

Writing and its subprocesses place a high demand on the storage and processing capacities of the working memory. The logic of speech recognition is to reduce cognitive demands, especially during the production of text, while increasing auditory resources available to aid rehearsal in a phonological loop (Kellogg, 1996). Quinlan (2004; 2006) has shown that less fluent writers show significantly increased text length and decreased surface errors during narrative creation by voice (speech recognition) as opposed to traditional text production by hand. Less fluent writers benefit from the lower physical effort in writing with speech recognition. The automaticity of text production is particularly important for skilled writing since general capacity may then be allocated to other subprocesses such as planning and revising (Bourdin & Fayol, 1994). It is not clear, however, if the execution characteristic of the writing mode is the most important characteristic writers benefit from[2]. For example, speech recognition generates only real words as errors because these items are part of the available lexicon while word processed errors can be typographical errors resulting in non-words.

[2] Most studies show that speech recognition could be less demanding to generate text (MacArthur, 2006; Quinlan, 2004; 2006) are done with special populations who already experience great demands from keyboard & mouse.

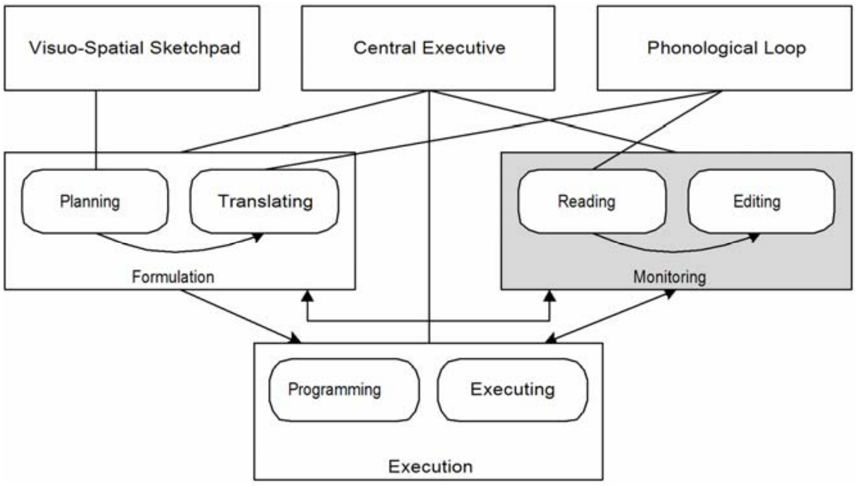


Figure 3. A representation of Kellogg's extension of Baddeley's model of working memory to writing. Adapted from Kellogg (1996).

In the vast amount of literature on working memory, several definitions have been mixed: working memory, cognitive load, cognitive resources, cognitive capacity and cognitive effort. As Figure 3 shows, the working memory exists of various components. Working memory refers to the ability of temporarily maintaining mental representations for the performance of a cognitive task. The cognitive load of a task refers to the load on working memory during problem solving activities (like writing). Piolat et al. (2004) present a clear distinction between the latter three definitions. Cognitive resources can be seen as the mental energy that is available at a certain moment during cognitive processes. The cognitive capacity is a more general measure that is quite constant per person. Each individual has a certain amount of resources that are available to him or her. This can be measured by various tests of working memory span (Daneman & Carpenter, 1980; Ransdell & Levy, 1999; Towse, Hitch, & Hutton, 1998). In this paper we are mainly interested in the cognitive effort a writing task imposes on a writer. This effort corresponds with the amount of resources that are required by a writing task. In this study we will gain insight in the cognitive effort it takes for writers to solve various error types during text production, while the TPSF is offered only visually or also via the auditory channel.

Writing with speech may alter balancing working memory resources relative to writing without speech. Since increasing numbers of novice writers are learning to use speech recognition in an effort to improve their writing – or who, in case of repetitive strain injury are forced to do so – it is imperative to understand the changes that are being caused by this new mode of writing. Furthermore, despite the differences related to the writing mode, we hope that these findings will also contribute to a better understanding of the interaction with the TPSF during writing in general, that is, when writing without speech recognition.

1.4 Hypotheses

We assume that working memory makes a substantial contribution to the primary task of error detection and text completion and that it competes for resources with the secondary task of responding to an auditory probe. By comparing error correction strategies we can then determine the relative contributions of cognitive effort from several sources, error span, input mode, and lexicality. We assume that the cognitive effort will differ for error types. By comparing error types we can determine the effect of error types on the working memory.

1.4.1 Effect of error presentation mode (speech vs. non-speech) on the cognitive effort of error correction

The first hypothesis compares the effect of the two main experimental conditions: the TPSF was presented either visually or was read aloud also via voice. We expect that errors that occur after the initial clause has been shown in the non-speech condition require less cognitive effort than errors that occur after the context is offered in the speech condition. This assumption is based on findings in research on classical dictation (Gould, 1978; Gould & Alfaro, 1984; Schilperoord, 1996). Writers who dictate their texts to a machine need to make a mental representation of the text. We assume that writers who only have to compare the TPSF to the expected content are

less distracted by the error than writers who have to compare the TPSF both to the expected content and the speech prompt that contained an error-free ‘voice representation’.

> Hypothesis 1: Errors that occur in the TPSF that were only showed visually cost less cognitive effort than errors that occur in the TPSF that also has been read aloud to the writers.

Our hypothesis is based on two lines of reasoning. As mentioned above, we expect that writers, who hear the dictated TPSF first, will be more focused on text production. In other words, offering the TPSF via speech might cause a focus on continuing text production. This positive characteristic of speech might have a counterpart. The focus on text production might cause an extra cognitive burden to switch to the revising subprocess and to evaluate the form and content of the TPSF. In other words, errors in the TPSF might distract writers because of the focus on text production. In theory writers do not need to read the TPSF, but the appearance of the TPFS on the screen – combined with positioning the cursor in the text – is such a strong trigger for text monitoring that we expect writers to ‘glance’ at the TPSF. We assume this action is rather superficial and mainly aimed at getting the ‘Gestalt’ of the text so as to continue production, rather than carefully rereading – and evaluating – the TPSF for further text production.

Secondly, we assume that in the auditory condition the writers may consider the reading of the TPSF as a kind of triple task: they get a written representation of the TPSF in the context screen, they hear the dictated form of the TPSF and then the (deficient) TPSF appears on screen. We expect that it is easier to detect and correct the error when writers only have to compare the visualization of the TPSF to the expected content and form. They will be less distracted by the error in the TPSF than writers who have to compare the TPSF on the basis of the (extra) auditory prompt that contains an errorless ‘spoken representation’ of the TPSF. Consequently, we expect it to be easier to detect an error in the TPSF in the condition without than with an auditory prompt.

1.4.2 Effect of error span on the cognitive effort of error correction

In the second hypothesis, we expect that large error spans (covering a character spread of more than two characters) cost more cognitive effort than small errors[3]. Large errors differ to a greater extent from the mental representation than smaller errors do. Leijten & Van Waes (2005) show that writers prefer to solve most large errors immediately after they appear/occur in the TPSF. We suggest two reasons. A first explanation can be that large errors are easy to detect and are therefore easy to solve immediately. Another explanation can be that large errors lead to a prominent deviation from the intended text that cannot be ignored. They impose a larger cognitive load on the writer and apparently create an urgent need to fine-tune the text again by correcting the error.

> Hypothesis 2: Large errors cost more cognitive effort than small errors.

Hypothesis 2 directly compares error spans. Error span refers to the number of characters separating components of an error. When the difference between the correct and the incorrect word is large (i.c. more than two characters), it may be easier to recognize the error, but at the same time, it may require more working memory resources due to the time delay required for maintaining the difference in representation. The time delay causes the need to re-read the sentence in order to engage in error correction or editing. This re-reading is known as highly complex and demanding on working memory (Just & Carpenter, 1992).

1.4.3 Effect of input mode on the cognitive effort of error correction

In the third hypothesis, we expect that small errors occurring in writing with keyboard and mouse take less cognitive effort to solve than small errors occurring in writing with speech recognition. Input mode refers to whether the error was naturally-occurring within text created on the computer by keyboard and mouse or speech recognition. Examples of typical speech recognition errors are possessive pronouns that become personal pronouns (mine vs. my) and double words if only one single word was intended (the the) and the insertion of full stops that unintentionally creates a new sentence (cf. section 2.3 Material).

> Hypothesis 3: Errors that are originated in keyboard based writing cost less cognitive effort than errors that are originated in speech recognition.

[3] For instance, the difference between the correct spelling of the word ‘speech recognition’ and the incorrect spelling of ‘speech recognitiion’ (small error) on the one hand, and of ‘speech regoingition’ (large error) on the other hand (cf. section Materials).

On average, writers have much more practice with the types of errors made with keyboards compared to those with a novel system like speech recognition. The present study will directly compare the specific effect of each input mode among.

We expect the physical environment of writing with keyboard & mouse to be more closely connected to the writer than the speech recognition environment. In this period of time, this closeness can be the decisive factor in detecting errors in the TPSF[4]. Hypothesis 3 will compare the effect of small errors caused by keyboard or speech recognition entry.

1.4.4 Effect of lexicality on the cognitive effort of error correction

In the fourth hypothesis, we refine the third hypothesis. We expect that small errors that occur only in writing with keyboard & mouse take less cognitive effort than small errors that can occur in writing with speech recognition or keyboard & mouse, because the writers in the previous mentioned case study (Leijten & Van Waes, 2005b) showed a unanimous preference to solve nonexistent words immediately.

> Hypothesis 4: Errors that are originated in keyboard based writing (non-existing words) cost less cognitive effort than errors that are originated in speech recognition or keyboard based writing (existing words).

In the normal course of events, non-existent words only occur in writing with keyboard & mouse. The speech recognizer will generate, by definition, only existing words.
Non-existing word: typing error in the word ‘streert’ instead of ‘street’.
Existing word: speech recognition error in the word ‘eye’ instead of ‘I’ (cf. section 2.3 Material).

Hypothesis 4 compares errors that involve lexicality (semantic level). Lexicality refers to whether the error is in a real word or a non-word. The former can be meaning-based or surface-based while the latter can only be surface-based and should therefore be easier to detect and correct. Real words should take greater resources to process than non-existing words, because the context involved in detecting the error necessarily exceeds the boundaries of the word itself. The findings in Hacker et al. (1994) and Larigauderie et al. (1998) suggest that spelling errors are easier to solve than semantic based errors, implying that non-existing word errors should be easier to solve than real word errors. In other words, we hypothesize that keyboard errors are easier to solve than errors that could occur as well in speech recognition as in keyboard based word processing.

In sum, the present study provides an analysis of the variations in cognitive effort related to error correction. The design includes the most frequently occurring error types found in a case study of professional writers (Leijten & Van Waes, 2003b, 2005b). The error types were presented to college students who were asked to detect errors and complete causal statements. An analysis of the task components of error correction will provide information about the mechanisms by which working memory resources constrain revision during writing.

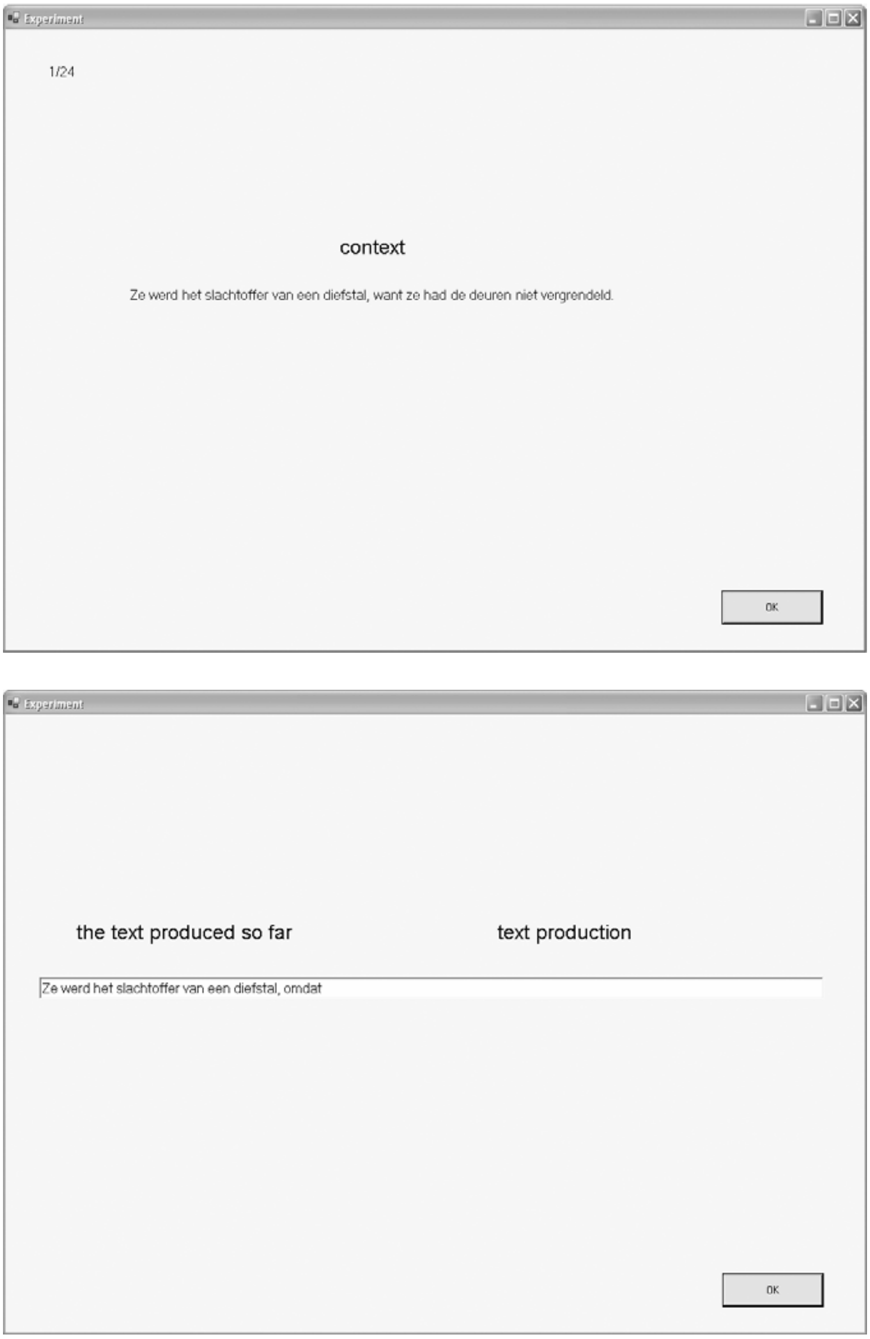
2 Method

To answer the research questions above, we set up a controlled reading-writing experiment. So far, error correction strategies using speech recognition have been described in natural writing tasks. In this study we opt to isolate various error types that are most common during writing with speech recognition and with keyboard & mouse. Furthermore, the complexity of the writing task is controlled for. In this study we compare many writers’ strategy decisions as a function of error type.

Participants were invited to participate in a one hour experiment during which they had to take two short initial tests and complete two sets of reading-writing tasks in two different modes, one purely visual task and the other a read aloud task before the visual representation of the TPSF appeared.

The task consisted of a set of sentences that were presented to the participants one by one to provide a new context. After every sentence the participants had to click the ‘ok’ button, indicating that they had finished reading the sentence.

[4] Please note that the errors mentioned here only occur in speech recognition based writing processes, but that the participants themselves do not use speech recognition during this study. The errors are just originated in the speech recognition mode.



Ze werd het slachtoffer van een diefstal, want ze had de deuren niet vergrendeld.

Ze werd het slachtoffer van een diefstal, omdat

Figure 4. Example of (annotated) context screen and text produced so far.

A subclause of the previous sentence was then presented as TPSF in another subordinate causal structure, and the participants were prompted to complete the sentence. Figure 4 shows an example of the context sentence (top) and the TPSF positioned in a box to complete the sentence (bottom).
Four kinds of errors were implemented in the TPSF. In this section, we shortly describe the profile of the participants, the design of the experiment, the software used, the materials that were used in the reading-writing tests, and the procedure the participants followed.

2.1 Participants

Sixty students participated in this experiment[5]. The students all had Dutch as their first language and were between 18 and 22 years old. They were randomly assigned to the experimental conditions. The participants who volunteered to take part in the experiment received a free movie ticket.

[5] In total, 67 students participated in the experiment. On the basis of technical aspects (missing values due to a technical problem in the logging file), and an outlier analysis of both the quality of the corrections (primary variable) and the reaction time behavior (secondary variable), 7 students were excluded from the study.

2.2 Design

The experiment employed a 2 (experimental condition: mode of presentation speech versus non-speech) by 2 (two sets of sentences) within-subjects design (see Table 1). We constructed two sets of sentences in which an equal number of errors was distributed in a comparable way. Error type was equally distributed. The order in which these sets of sentences were presented to the participants was counterbalanced in the design of the experiment. The sequence of how the sentences were offered was also varied (only visual/non-speech or read-aloud and visual/speech).

Table 1. Experiment design

Number of group	Order of experimental condition (speech, non-speech and set of sentences)	
Group 1	non-speech | set of sentences 1	speech | set of sentences 2
Group 2	non-speech | set of sentences 2	speech | set of sentences 1
Group 3	speech | set of sentences 2	non-speech | set of sentences 1
Group 4	speech | set of sentences 1	non-speech | set of sentences 2

2.3 Materials

The main part of the experiment consisted of a reading-writing task. The participants had to read and complete 60 short sentences. They first read a short sentence which provided a context for the next sentence that had to be completed in the next step of the procedure.

Example:
Context: Because it has rained, the street is wet.
Correct TPSF: The street is wet, because …
Incorrect TPSF:The streert is wet, because …

All experimental sentences that contained deficiencies were marked by a causal coherence relationship. All the materials were presented in Dutch.

In 24 out of the 60 sentences that were used in the experiment, we varied four types of errors to construct the deficient TPSF based on error span, input mode and error lexicality. The errors were taken from a larger corpus of data collected in a previous study on the influence of writing business texts with speech recognition (cf. section 1.4). The types of errors were replicated in the sentences built for the current experiment. The errors we selected were either caused by writing with speech recognition or by writing with keyboard and mouse.

An example of a typical error in the speech recognition mode[6] caused by a misrecognition of the dictated text is:

Example:
(3) Spoken input 'The street is wet, because it has rained.'
(4) Incorrect output 'The street is wet, because it has **drained**.'

This kind of error will not occur in writing with keyboard & mouse. Other mistakes however could be classified as 'mode independent':

Example:
(5) 'The **streert** is wet, because it has rained.'

[6] Speech recognition always produces correct words. The words can be incorrect in a certain context, but the words themselves exist and they are always spelled correctly.

Because the ‘d’ and ‘r’ are adjacent keys on most keyboards, a writer could make this type of error easily. The typing error in this example leads to another existing word. Therefore, this error could also occur in the speech recognition mode. So although, the process that leads to the error may be different (ergonomic versus phonological), the written representation can be identical. On the other hand, some type of errors will not occur in speech recognition, and are exclusively found in texts produced with keyboard and mouse. These kinds of errors result in non-existing words.

Related to these characteristics of errors occurring in speech recognition and in writing with keyboard and mouse, we also decided to differentiate the size of difference (number of characters that are different between the intended word (clause) and the actual representation). Table 2 shows the classification of the errors taken into account the mode-specific characteristics of speech recognition and keyboard based word processing. An example of the four error types as they are included in the experiment can be found in Appendix 1.

Table 2. Classification of errors

	Type of error		
Category	**Error span**	**Input mode**	**Lexicality**
SR Large	large: > 2 characters	only in speech recognition	existing words
SR Small	small: ≤ 2 characters	only in speech recognition	existing words
SR Keyboard | Small	small: ≤ 2 characters	in speech recognition and	existing words
Keyboard Small	small: ≤ 2 characters	only in keyboard & mouse	non-existing words

The location of the errors in the sentences also varied. In each category, half of the errors were placed at the beginning and half of the errors at the end of the first text segment. One sentence was also constructed in each category in which there was an error at both locations. These sentences were randomly distributed in each test set. Finally, so as not to create a default attitude that was preconditioned to finding errors in the TPSF clauses, more than half of the sentences that were presented were correct and contained no mistakes. Some of these so-called filler sentences were constructed with a temporal relation instead of a causal structure. There was a concentration of correct sentences at the beginning of each set.

Example:
(1) Spoken input ‘I am writing a short text.’
(2) Incorrect output ‘**Eye** am writing a short text.’

Four of the correct sentences were ‘mirror’ sentences of the incorrect sentences, allowing us to compare the reading-writing behavior in an even more controlled way on a small set of sentences. Table 3 gives an overview of the characteristics of each set of 30 sentences.

Table 3. Overview of the characteristics of the sentences used in each reading-writing task

Characteristics	Number of sentences
Test sentences (4 correct, 2 incorrect)	6
Incorrect sentences (4 type of errors * 3 places of occurrences)	12
Correct sentences	8
Correct sentences based on incorrect sentences (one sentence for every type of error)	4
Total	30

All sentences were more or less equal in length: the contexts contained 70 to 90 characters; subordinate clauses 31 to 43 characters; number of characters context set 1, *M* = 78.42 and set 2, *M* = 79.22.

Example:

Context
She was not allowed to leave the hospital, because a complication had arisen.
(Dutch: Ze mocht het ziekenhuis niet verlaten, want er trad een complicatie op.)

Text segment with error
She was not allowwed to leave the hospital, because…
(Dutch: Ze mochtt het ziekenhuis niet verlaten, omdat…)

extremely serious   not so serious
□ □ □ □ □ ■ □

Before the materials were used in the experiment, the manipulation of the sentences was pretested on validity among 32 students (17 male, 15 female). The students were asked to read 29 sentences consisting of a context and a short text segment that included information from the context (see example below).

Each sentence contained one or more errors. The errors were all marked. For each error the participants had to assess the seriousness of the error on a 7-point Likert scale. They were instructed to read the sentences as if they were part of a first draft of a group work. They had to indicate the seriousness of the error.

Context
Because it rained every day, the year 1998 is a bad wine year.
(Dutch: Omdat het elke dag geregend heeft, is het jaar 1998 een slecht wijnjaar.)

Text segment with error
The year conscription 98 is a bad wine year, because…)
(Dutch: Het jaar lichting 98 is een slecht wijnjaar, want…)

A cluster analysis was used to verify whether the sets of sentences were manipulated and classified correctly. Sentences that did not fit within a cluster that corresponded to the predefined category were deleted from the material. For instance, a sentence that was categorized as an error with a large error span and was assessed and clustered as a not so serious error, was deleted from the set. For example:

The hierarchical cluster analysis showed that this sentence was clustered with small errors, although we had technically classified it in the category of large errors.

2.4 Procedure

The participants were assigned to groups of four to participate in the experiment. In the computer laboratory, they were separated from each other by large partitions that were put between the rows of tables in order to avoid visual distraction. Before the participants started with the experiment, they had to put on a headset and position the button for the reaction test on the side of their non-dominant hand. Then they had to fill out some general information about their profile (name, age, education) on an on-line form. The researcher assigned a group number (1 to 4) to each of the participants. Next, a general overview of the experiment was provided to the participants that they could both read on the computer screen and listen to through their headset because the texts were also read out loud.

The experimental session consisted of five parts:

- Counting Span Test
- Baseline Reaction Time
- First part reading-writing test
- Second part reading-writing test
- Questionnaire

Before a new part started, the participants systematically received a written and an oral introduction to the new task (the instructions can be found in Appendix 2). A short trial task preceded every main task. To manage the experiment and the different flows, a special program was developed (Microsoft Visual Basic.Net). The program controlled the design, and stored the results of the tests, as well as several time stamps, the text produced in the writing task and the text location. We also integrated a Counting Span Test (CSTC) by Levy & Ransdell (2001)[7] to assess the participants' cognitive capacity. To log the linear development of the writing process during the completion task, Inputlog (Van Waes & Leijten, 2006) was used to capture the keyboard & mouse input and calculate the pausing time afterwards.

The first test was the *Counting Span Test*. This method assesses how well people can store and process information at the same time (working memory span). The Counting Span Test involves counting simple arrays of objects while simultaneously trying to remember an ever-increasing number of counts that are required as the memory sets become larger (Adams, Hitch, & Hutton, 1997; Levy & Ransdell, 2001; Ransdell & Hecht, 2003; Ransdell & Levy, 1999; Towse, Hitch & Hutton, 1998)[8].

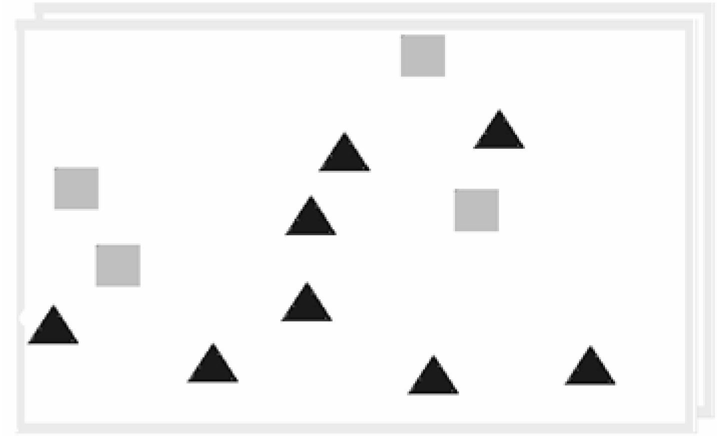


Figure 5. Example of card for Counting Span Test: squares to count (grey) and triangles to ignore (black).

The participants were instructed to count all the blue squares on cards with blue squares and red triangles (see Figure 5). As soon as the participants had finished counting the blue squares, they had to type in the correct number. Then another card appeared and again the correct number of squares had to be typed in as fast as possible. After two, three or four cards, small blank textboxes were presented on the screen and the participants had to sequentially recall the number of squares on the previous cards.

The Counting Span Test started with two count cards as a test. After this trial session, the participants had to go through two main sessions. In both sessions the complexity of the task was gradually increased. The general measurement of working memory span was quite equally distributed over the participants. Almost all the participants reached the highest level of the Counting Span Test. As we expected the cognitive capacity of participants did not seem to differ very much. In Table 4 the general data of the Counting Span Test are shown.

Table 4. General data Counting Span Test

		Percentage of correct items		Mean time in seconds	
	N	*%*	*SD*	*M (s)*	*SD*
Level 2	52	95.51	19.83	4.32	1.33
Level 3	50	95.33	16.51	5.47	1.41
Level 4	50	91.84	22.08	7.31	2.06

The second test was aimed at measuring the mean baseline *interference reaction time* of the participants. As stated above, one of the most powerful ways of discovering working memory contributions to writing has been to employ dual-task techniques (Baddeley & Hitch, 1974; Kellogg, 1996, 2001; Levy & Ransdell, 2001; Olive, 2004;

[7] This program can be obtained from www.psychologysoftware.com.

[8] Towse et al. (1998) are cautious about this method, but Ransdell & Hecht (2003) showed cross-age effects. To increase difficulty we have chosen to perform the large card final method (the large card last condition required holding the count over a longer retention interval).

Ransdell, Levy, & Kellogg, 2002). Fisk, Derrick and Schneider (1986-87) review the criteria by which such techniques are most valid. First, the primary and secondary tasks must require a trade-off between common resources needed for processing and storing information. A vast literature on working span relies on a dual task in which words, numbers, sentences, or other stimuli are processed for meaning while other information must be stored for later processing (Chenoweth & Hayes, 2003; Daneman & Carpenter, 1980; Hoskyn & Swanson, 2003; Kellogg, 1999; Ransdell & Hecht, 2003; Ransdell & Levy, 1999). Second, the primary task performance must decrease as a result of the addition of the secondary task. In writing the degradation of writing performance is seen as evidence that the primary task writing suffers from the secondary task: cognitive resources overlap. An opposite effect can be interpreted in the same way. Longer reaction times to a secondary task can be interpreted as a high cognitive effort that needs to be invested in the writing task at the moment of the secondary task (Olive & Kellogg, 2002). Third, the resources allocated to each task must be stable. Implications for error analysis during writing mean that: the primary task, error detection and completion, must compete with the secondary task, in this case, responding to a variable interval tone; the primary task must be overloaded by the secondary; and there must be practice to reach a consistent level of performance. When conducting a secondary task it is also necessary to measure the mean baseline reaction time as a reference measure. Thirty auditory probes were randomly distributed in an interval with a mean of 8 seconds and a range of 2 to 12 seconds. Participants were asked to press a button as rapidly as possible whenever they heard an auditory probe. After every probe the participants were asked to reposition their hands on the keyboard. The median baseline reaction time of each participant was calculated from the 25 last reaction times. The first five probes were treated as trial probes.

The third and the fourth tests were *reading-writing tests* with or without the addition of a spoken script prior to the visual presentation of the clause (TPSF). The participants were also informed that during the writing tests they would occasionally hear auditory probes (beep tones). They were asked to react as rapidly as possible to these probes by pressing the special button. During the reading-writing tests, the probes were distributed semi-randomly over the sentences that had to be completed. In the sentences with an error the probes were always presented in such a way that they occurred during the reading process; in the other sentences, especially in the test phase and in the non-causal sentences (temporal sentences), they were randomly distributed either in the reading or the writing phase. In some sentences the probe was not offered so as not to condition the participants.

The participants were informed that they should complete the sentences and that they should always try to write correct sentences. They were also told that they should focus both on accuracy and on speed. They had to finish the sentence as fast as possible and they had to – if necessary – correct the errors in the part of the sentence that was presented as a TPSF prompt. It was also explicitly mentioned that they should decide themselves if they preferred either to correct the sentence first, or to complete the sentence first and then correct the TPSF, if necessary. Next to this they were also instructed to respond as rapidly as possible to the auditory probes.

In the final *questionnaire* the participants were asked about their previous experience in computer-based word processing and speech recognition. Additionally they were asked about the task complexity and to indicate which sentences they did not fully complete and why.

2.5 Dependent variables

In order to analyze cognitive effort and error correction strategy, six dependent variables were derived from the logging data of the TPSF-program and Inputlog.

(a) Preparation time

The preparation time was defined as the time that passed between the moment the context screen was closed and the first mouse click to position the cursor in the TPSF screen, either to complete the sentence, or to correct an error in the TPSF. Since the items in the speech condition appeared only after the spoken script was presented, the duration of this script was subtracted from the preparation time.

(b) Delayed error correction

For every sentence we logged whether the cursor was initially either positioned within the TPSF clause that was presented as a writing prompt, or after the clause. We used these data as an indication of the participants' preference to correct the error in the TPSF first or to complete the missing part of the sentence first (and delay the correction of the error, if at all it was noticed). Inputlog enabled us to code these data. The initial position of the cursor in the text completion part was programmed as a set x-value.

This corresponds to a fixed x-value in the logging data, that is, 380: lower values refer to a cursor positioning in the TPSF and vice versa.

```
Example of logging output:
(1) Left Button: 00:00.15,109 (ClockTime) 274 (x-value) 345 (y-value) = in TPSF-
segment.
(2) Left Button: 00:00.18,329 (ClockTime) 380 (x-value) 339 (y-value) = in text
completion segment.
```

In our experiment the participants preferred to complete the sentence first in 89.2% (*SD* = 15.6) of the items. The individual preference ranged between 50% and 100%.

(c) Production time (writing)

The production time was defined as the period between the moment when the screen with the context sentence was closed and the moment when the screen with the TPSF to be completed was closed. In this period the TPSF was read, the sentence was completed, and possible errors in the TPSF were corrected. For the items in the speech condition, the duration of the spoken script was subtracted from the production time.

(d) Production quantity

In the analysis of the production quantity, the number of characters that were generated in completing the TPSF was calculated. The production quantity in the revision of the first clause of the sentence was not included in the analysis. We also calculated the difference between the correct completion of the sentence (theoretical number of characters to be produced based on the context) and the actual number of production quantity by the participant in the experiment. On average 30 characters had to be produced (about 5 to 6 words per clause). The number of characters to be produced in every item was kept almost constant for every item (*SD* = 0.7). Table 5 shows that the actual number of production quantity by the participants highly corresponded to the theoretical optimum: the average difference is only 1.3 characters.

The high equality between the theoretical optimum and the experimental realization showed that the writers succeeded in the text completion task: they completed the content of the text as requested (error correction is measured in accuracy, cf. below).

Table 5. Mean number of characters in completing the TPSF-clause

	Mean number of characters	
	M	*SD*
Theoretical optimum	30.7	0.7
Experimental realization	29.4	1.8
Difference	1.3	1.0

(e) Accuracy (quality of the correction of errors)

Accuracy here represents the percentage of sentences with a (manipulated) error that were rewritten correctly. Half of the sentences in the experiment contained either one or two errors which were always presented in the first clause of the sentence that had to be completed (TPSF). The participants corrected an average of 86.6 percent of the errors (*SD* = 8.4). In the analyses we only integrated the sentences that were corrected.

(f) Reaction time

The reaction time was defined as the time that passed between the moment when the auditory probe (beep) was given and the moment the button was pressed. The median reaction time of the baseline test for each participant (cf. initial reaction time test) was subtracted from the reaction time of each item to correct for individual differences. When the participant failed to press the button before completing the sentence, we coded the reaction time as missing; when the reaction was shorter than the participant's median of the baseline test, the reaction time was recoded to a zero value (representing a very short reaction time). Because we wanted to be sure that the reaction time in incorrect sentences was strictly related to the error and that the error was noticed as such, the reaction time for the incorrect sentences was coded as missing in those cases where the TPSF was not revised.

3 Analyses

To test the hypothesis we used the General Linear Model for the overall analyses, in which we compared the four error types and the experimental condition. To test the individual hypothesis per experimental condition (speech versus non-speech) we used the paired samples t-test[9], because of the within-subjects design.

The analyses are based on the causal sentences that contained one error and that were corrected by the participants. We assume that errors that are not corrected are not noticed by the writers. The behavior of the participants is then the same as for the correct sentences. That is why we have only taken into account the corrected sentences.

4 Results

A set of six variables is used to test our hypotheses. Before we present the results of the statistical tests related to the hypotheses, we briefly report the general findings comparing the participants' interaction with the correct and the incorrect TPSF (overall and mirror analysis).

4.1 General findings

These results create a framework for the more specific error analyses presented in the results section. For the correct items, we excluded the sentences with a temporal relation as they were included in the experimental set as distracters. We only included the results for the correct sentences with a causal relation between the TPSF clause and the missing clause, as is the case in all the incorrect sentences. For the incorrect items, only those sentences that contained a single error in the TPSF were included.

Table 6 shows that when a correct TPSF clause is presented, the participants need less time before they decide to start completing the sentence ($t(59) = -5.25$, $p < .001$). This small difference is not confirmed in the mirror analysis in which we limited the analysis to the eight items that were both presented correctly and incorrectly (cf. 2.3 Materials).

Table 6. Mean values for the correct versus the incorrect sentences

	Correct		Incorrect		
	M	*SD*	*M*	*SD*	*p*
Overall analysis					
Preparation time (s)	1.46	0.51	1.70	0.60	**1
Delayed error correction (%)	98.43	6.15	80.31	2.16	**1
Production time (s)	14.34	3.46	17.17	4.35	**1
Reaction time (ms)	250	125	272	142	*2

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

[1] The significance is based on the overall data: GLM for the two conditions on the one hand and the four error types.

[2] The significance is based on the overall data: t-test for the two conditions (median).

The overall analysis of the delayed error correction at the start of the completion of the TPSF indicates that for about 20% of the incorrect items the participants preferred to correct the error prior to the completion of the sentence[10].

The production time needed to complete (and correct) the TPSF is shorter for the correct items (M = 14.34 seconds, SD = 3.46) as opposed to the incorrect items (M = 17.17 seconds, SD = 4.35, $t(59) = -10.956$, $p < .001$). The analysis of the mirror sentences confirms this finding. The extra production time is partly used to correct the error, but in general also the time needed to read an incorrect TPSF takes significantly longer (cf. preparation time).

[9] The reaction times are per hypothesis based on the mean values. However, the overall measurement is performed with the median reaction times, because the largeness of the overall data we calculated the median here to reduce the impact of outliers.

[10] We should take into account that the analysis of the initial cursor position is only an indication of the correction behavior. Of course, a participant can initially decide to complete (part of) the sentence, positions the cursor in that part, but then changes his mind and corrects the error first anyway. The deviation of the 100% score for the correct sentences in the general analysis is to be interpreted parallel to the noise factor in the incorrect items.

The overall analysis also reveals a significant difference in reaction time ($t(59) = -2.284$, $p < .05$). As expected, the cognitive effort is affected by the opposition correct versus incorrect TPSF.

4.2 The effect of mode of error presentation (speech vs. non-speech)

The first hypothesis about the experimental condition (speech versus non-speech) was not confirmed by the preparation times, the strategy inferred from delayed error correction, nor by the reaction time. Errors that occurred in the TPSF after the context had been prompted visually were solved differently than errors that occur after the context is offered both visually and auditory. Delayed error correction showed that the writers delayed error correction in the TPSF more often when the TPSF was presented also in speech. The results of the preparation time analysis (Table 7) build on this result.

Table 7. Mean values in both the speech and non-speech conditions

	Speech		Non-speech		
	M	*SD*	*M*	*SD*	*p*
Preparation time (s)	1.10	0.47	2.38	1.22	**1
Delayed error correction (%)	93.81	17.63	66.07	35.42	**1
Production time (s)	17.86	4.72	18.89	5.38	-1
Accuracy (%)	87.85	11.02	88.11	8.64	-1
Reaction time (ms)	254	158	284	166	*2

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

[1] The significance is based on the overall data: GLM for the two conditions on the one hand and the four error types.

[2] The significance is based on the overall data: T-test for the two conditions (median).

The speech condition of writing led to reliably faster preparation times and delayed error correction indicating that the speech condition leads to a more explicit preference to complete writing first and correct the error only after writing. The faster preparation time indicates that writers more superficially read the TPSF in the speech condition, which is in line with the shorter reaction time during reading in this condition. No other measures showed a significant difference between both conditions.

The preparation time for the items with an error in the TPSF in the condition without speech is longer than with speech (speech: $M = 1.10$, $SD = 0.47$ vs. non-speech: $M = 2.44$, $SD = 1.22$), $F(1, 59) = 79.576$, $p < .001$). The results for the analysis of the delayed error correction reinforce this finding and reveal another interaction with the TPSF: in the speech condition, the participants tended to complete the sentence first, while in the non-speech condition, they tended to correct the error first (speech: $M = 93.81$, $SD = 17.63$ vs. non-speech: $M = 66.07$, $SD = 35.24$), $F(1, 59) = 26.716$, $p < .001$). However, the results of the accuracy test show that the participants were not more successful in the speech condition than in the non-speech condition (speech: $M = 87.85$, $SD = 11.02$ vs. non-speech: $M = 88.11$, $SD = 8.64$). Apparently, participants only preferred to correct the errors after having completed the sentence. On the basis of the data one cannot make a quality inference about the difference between a strategy of ‘ignoring’ or ‘overlooking’ the error in the first phase of completing the TPSF. There was a significant effect found in the cognitive effort: the reaction time in both modes was significantly different (speech: $M = 254$, $SD = 158$ vs. non-speech: $M = 284$, $SD = 166$); $t(59) = 2.47$, $p < .05$). In general, the speech condition freed resources to respond faster to the secondary task.

Apparently, writers are not aware of this positive effect of speech recognition. Half of the participants indicated in the questionnaire that the writing task without the addition of the auditory information was the easiest (46.55%), and the other half indicated that this addition made the task easier (53.45%). Only 8% of the participants had previous experience in speech recognition, only one of them had more than 20 hours of practice. They used it for word processing and for chatting.

4.3 The effect of error span on cognitive effort

Table 8 shows that the preparation time for the items with a large error in the TPSF is longer than that for the small errors (SR Large: $M = 2.07$, $SD = 0.47$ vs. SR Small: $M = 1.48$, $SD = 0.54$, $F(1, 59) = 36.39$, $p < .001$). The interaction between the type of error and the speech condition is also significant ($F(1, 59) = 13.37$, $p < .01$). Larger errors take more time to process, especially when the text is not preceded by speech. The main effect of delayed error correction was non-significant (*F*(1, 59) = 2.57, *p* = .115). However, the interaction effect was significant (*F*(1, 59) = 5.53, *p* < .05), indicating that writers in the non-speech condition preferred to correct the large error first and after that complete the sentence. The T-test confirms this finding (*t*(59) = -2.38, *p* < .05).

Table 8. Mean values in both the speech and non-speech condition for error span

	SR Large Error		SR Small Error		
	M	*SD*	*M*	*SD*	*p*
Preparation time (s)	2.07	0.94	1.48	0.54	**
Speech	1.20	0.70	1.00	0.49	*
Non-speech	2.95	1.67	1.97	0.86	**
Delayed error correction (%)	78.75	25.14	82.92	24.13	-
Speech	93.33	21.52	92.50	24.05	-
Non-speech	64.17	41.26	73.33	37.36	*
Production time (s)	20.48	6.40	15.61	4.39	**
Speech	20.83	7.97	15.32	5.05	**
Non-speech	22.12	8.22	15.89	5.25	**
Accuracy (%)	92.50	16.14	84.17	19.51	*
Speech	92.50	22.22	85.00	26.52	-
Non-speech	92.50	20.22	83.34	23.77	*
Reaction time (ms)	372	255	303	161	*
Speech	332	344	270	172	-
Non-speech	423	309	345	233	**

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

In both conditions, the writers needed more time to produce the text segment with the large error (*F*(1, 59) = 115.95, *p* < .001). The type of mode does not affect this finding. A closer look at the accuracy of solving the errors in the text correctly shows that large errors, those that take more time to process, are easier to correct. The accuracy for the large error is higher than for small errors (*t*(59) = 2.38, *p* < .05). Smaller errors are more easily overlooked, especially in the mode of writing without speech (*t*(59) = 2.10, *p* < .05). Whereas smaller errors are easier to overlook, larger errors distract more than smaller errors. The reaction time on the auditory probe is larger, and therefore takes more time, in sentences with a large error (*F*(1, 59) = 4.35, *p* < .05).

In our hypothesis we stated that it may be easier to recognize larger errors, but at the same time, it may require more working memory resources due to the time delay required for maintaining the difference in representation. To find evidence for this statement we analyzed the pausing behavior before the error correction. The pausing time before a revision in the TPSF is an important indicator of the cognitive effort it takes to correct the error. For the coding of this variable we used the linear text representation generated by Inputlog. This XML file enabled us to identify each pause before the actual error correction. The pause threshold value for the analysis was 500 milliseconds. We coded initial error correction in TPSF (=immediate error correction) and error correction that followed text production (=delayed error correction). If writers start with error correction than the pausing time is equal to preparation time (positioning of cursor near error in TPSF). In cases where revision follows text production the pause is measured from the end of the text production until the movement of the cursor to the error in the TPSF.

Example:
Context: Because the topic interests me, I will go to the conference next week.
(Dutch: Omdat het thema mij interesseert, ga ik volgende week naar die conferentie.)
TPSF incorrect: I am going to the **competence** next week, because ….
(Dutch: Ik ga volgende week naar die **concurrentie**, want…)

Figure 6 shows an example of both error correction types. The writers are both completing the next sentence (literal translation from Dutch - not phonological - for the error):

time	linear representation	explanation
14:20,00- 14:59,00	Left Button[947,710]	click OK ([x,y] within zone 'ok'-button)
	Movement[340,342]	positioning in TPSF-zone
	{4000} Left Button[340,342]	pause of 4 sec before revision (=immediate)
	Movement[454,350] {672}	
	BS 8	deletion by backspace of 'concurrentie'
	{688}fer{766}entie{547}	typing of 'ferentie'
	Movement[414,352]	mouse movement to production-zone
	{844}het ·thema · interesseert ·mij.	finishing sentence
	Movement[904,697]	
	{1141} Left Button[904,697]	click OK

time	linear representation	explanation
14:20,00- 14:59,00	Left Button[893,711]	click OK
	{3266}	pause of something over 3 seconds
	Movement[388,345] {547}	positioning in production-zone
	Left Button[389,345]	
	{687}het ·thema · inte{672}resseeer BS BS rt ·mij.	sentence completion (correction typing error)
	{1203}	pause of 1.2 sec before revision (=delayed)
	Movement[280,346] Left Button[295,345]	mouse movement to TPSF-zone selection of 'cur'
	Movement[308,350] fe	
	Movement[912,702] {3328}	replace by 'fe' (resulting in conferentie)
	Left Button[912,702]	
		click OK

Figure 6. Example of immediate versus delayed error correction.

Results in Table 9 show that writers paused significantly longer before a larger error then before a small error (*F*(1,17) = .359, *p* < .01). This result is the same for the speech and non-speech condition.

Table 9. Mean pause time for immediate and delayed correction for error span

	SR Large Error		SR Small Error		
	M	*SD*	*M*	*SD*	*p*
Pause time (ms)	2.27	1.32	1.64	1.16	**
Immediate	4.72	2.34	2.93	0.88	**
Delayed	0.92	0.42	1.07	0.70	**

** *p* < .01

There is a strong interaction effect between the pausing behavior of error span and immediate and delayed error correction (*F*(1,17) = .350, *p* < .01). The pausing times before errors that are corrected immediately is probably distributed over several processes: reading the TPSF, detecting the error and perhaps building a mental representation of the text completion (not necessary in this order). This may account for the longer initial pause durations. When error correction is delayed, the focus is most likely on error detection, diagnosing and correction. The example in Figure 7 shows that the error detection has already happened during the initial preparation time, but that error correction is delayed. In those cases writers prefer to continue with text production first, probably to free their cognitive resources from maintaining the context information in their short term memory.

The example shows that the error detection probably took place during the initial pause of the preparation time. This explains why a pause of half a second is sufficient to position the cursor near the error in the TPSF. Other data show an even shorter pause before this positioning. The diagnosis of the error and the determination of the correction strategy (e.g. deleting the whole word versus opting for letter substitution) take more time and elapse in a more fragmented manner.

linear representation	explanation
Left Button[905,706] {2891} Movement[437,340] {563} het ·t{750}hema ·interesseert · mij. {531} Movement[286,348] {2156} Left Button[293,348] {1141} Movement[303,346] {703} BS BS BS fe{672} Movement[929,707] {1000} Left Button[928,706]	click OK initial pause of almost 3 sec and positioning in productionzone sentence completion (pause in 'thema') short pause before movement to TPSF pause of 2 sec and positioning of cursor pause of 1 sec and small mouse movement short pause and change of 'concurrentie' in 'conferentie' pause of 1 sec and click OK

Figure 7. Example of delayed error correction.

The pausing data might provide more information about the moment the participants detect and diagnose the error. The pausing data show various possible profiles. Short initial pauses lead to a fast decision process by the writers. Writers decide to correct the error immediately or continue text production based on the gist of the text. During longer initial pauses various strategies might be conducted. Writers could have long initial pauses that lead to immediate error correction. During the initial pause the writer focuses on sentence completion. He re-reads the TPSF in order to complete the text (grasping the gist of the text for further text production) and is confronted and consequently distracted by the error in the TPSF, which leads to immediate error correction. Conversely, the long initial pause could lead to delayed error correction. During the initial pause the writer again re-reads the TPSF in order to complete the text, and is confronted with the error in the TPSF. However, he decides to prioritize sentence completion and to delay error correction. The error detection is stored in short term memory to help locate the error after sentence completion.

The latter strategy can cause information loss. The example in Figure 8 provides an example of this situation. We are aware of the fact that we do not have direct evidence of the eye movements of the participants. This additional data could confirm our assumptions. The example is therefore more illustrative (cf. discussion). The example shows a rather long initial pause and a delay of error correction. The writer seems to hesitate on the word 'interest' and a typing error distracts the writer during text production, which again causes a longer (re-)orientation on the TPSF.

The examples in Figure 7 and Figure 8 show that writers have several options when allocating their cognitive resources in re-reading the TPSF and correcting the error. They also show that extra time delay does cause the need to re-read the sentence in order to engage in error correction. The data show that re-reading indeed is demanding on cognitive resources.

Section 4.4 and 4.5 provide more information on the error types, input mode and lexicality.

linear representation	explanation
Button[907,693] Movement[393,346] {3468} Left Button[393,345] {687} het ·thema{796} · {907}inter{594}esser {640} BS et BS rt ·mij {3281}.{859} Movement[297,348] {2515} Left Button[296,348] Left Button[297,348] {1094} conferentie {1531} Movement[916,702] Left Button [916,700]	click OK and direct movement to production zone, initial pause of 3.5 sec and repositioning sentence completion (pause between and in words) short pause, double correction of typing error and completion sentence pause before and after full stop and positioning in TPSF, pause of 2.5 sec selection of TPSF-error (full word by double click) pause of 1 sec and error correction pause of 1.5 sec click OK

Figure 8. Example of delayed error correction (with additional error correction).

4.4 The effect of input mode on cognitive effort

The results in Table 10 show that the participants were equally distracted by the typical small speech recognition errors (cf. (d)rained) and the keyboard errors. The error categories do not lead to a difference in preparation time (*F*(1, 59) = 2.854, *p* = .096). However, the results show that the preparation behavior, especially when confronted with small speech recognition errors, is very diverse among the participants (Keyboard Small: *M* = 1.20, *SD* = 1.21). The production time and the reaction time are comparable for both error types.

Table 10. Mean values for small errors in both the speech and non-speech condition for the relation between the input mode and the error

	SR Small Error		Keyboard Small Error		
	M	SD	M	SD	*p*
Preparation time (s)	1.48	0.45	1.62	0.76	-
Speech	0.99	0.49	1.20	1.21	-
Non-speech	1.97	0.86	2.03	0.90	-
Delayed error correction (%)	82.92	24.13	77.08	24.91	**
Speech	92.50	24.05	92.50	24.05	-
Non-speech	73.33	37.36	61.66	43.54	*
Production time (s)	15.61	4.39	15.70	4.46	-
Speech	15.32	5.05	15.33	4.58	-
Non-speech	15.89	5.25	16.06	5.79	-
Accuracy (%)	84.17	19.51	96.25	9.00	**
Speech	85.00	26.52	95.83	13.94	**
Non-speech	83.33	23.77	96.67	12.58	**
Reaction time (ms)	313	171	332	201	-
Speech	286	196	328	239	-
Non-speech	342	232	344	242	-

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

Although the participants do not seem to be more distracted by one error category or the other, they do show a stronger tendency to immediately correct the typical keyboard error before completing the TPSF. The typical speech recognition errors do not seem to initiate this kind of behavior (*F*(1, 59) = 8.48, *p* < .01). Especially if the typical keyboard error is not preceded by speech, the participants prefer to solve the error first (interaction effect: *F*(1, 59) = 8.48, *p* < .05).

Our expectation that small keyboard and mouse writing errors can be better solved than small speech recognition writing errors is confirmed by the accuracy of both error types. The typical keyboard errors seem to be easier to solve, because the accuracy of typical keyboard errors is higher than the accuracy of the speech recognition errors (SR Small: *M* = 84.17, *SD* = 19.51; Keyboard Small: *M* = 96.25, *SD* = 9.0 (*t*(59) = -4.66, *p* < .001). In both conditions, that is, speech and non-speech, typical keyboard errors have a higher accuracy (speech: *t*(59) = -3.423, *p* < .001, non-speech: *t*(59) = -4.000, *p* < .001).

4.5 The effect of lexicality on cognitive effort

One of the characteristics of using speech technology as a dictating device is that errors that are generated through misrecognition, always result in existing, correctly spelled words or word groups (‘eye’ instead of ‘I’). This is in contrast with most errors that occur in texts produced by keyboard & mouse. A typing error is often caused by pressing the key of an adjacent letter on the keyboard or just missing a key. Generally these typing errors result in non-existing words (‘streert’ instead of ‘street’). It was expected that small typing errors resulting in non-existent words (Keyboard Small) are easier to identify when reading the TPSF and can be solved more efficiently, and with less cognitive effort than small errors that result in existing words (SR | Keyboard Small).

Most of the results shown in Table 11 do not support this hypothesis. Neither the preparation time nor the production or reaction time seems to be influenced by lexicality. Based on these analyses, the interaction with the TPSF is not influenced by the fact that the error in the TPSF is an existing word or not. However, in the latter situation, the participants prefer to complete the sentence first before correcting the error in the TPSF, especially if the TPSF clause is not read aloud. This is similar to findings in our previous study with novice speech recognition users who also did not tolerate non-words and corrected them immediately. (SR | Keyboard Small: *M* = 70.83%, *SD* = 39.37 vs. Keyboard Small: *M* = 61.66%, *SD* = 43.54, *F*(1, 59) = 5.506, *p* = .022 (no interaction effect: *p* > .05). Of the sixty participants, 17 always correct the error in the TPSF first for the items of the non-lexical error category (Keyboard Small) whereas 12 are characterized by a mixed pattern, explaining the high standard deviation for this variable.

The only variable that supports the hypothesis about lexicality is the accuracy (Table 11). The analysis of the quality of the error correction yielded a main effect of error category (Existing Words SR | Keyboard Small: *M* = 89.17%, *SD* = 21.33 vs. Non-existing Words Keyboard Small: *M* = 96.25, *SD* = 9.00; *t*(59)= -3.752, *p* < .001).

Table 11. Mean values in both the speech and non-speech condition for lexicality of the error

	SR | Keyboard Small Error Existing Words		Keyboard Small Error Non-existing Words		
	M	*SD*	*M*	*SD*	*p*
Preparation time (sec)	1.62	0.72	1.62	0.76	-
Speech	1.52	0.69	1.20	1.21	-
Non-speech	2.08	1.10	2.03	0.90	-
Delayed error correction (%)	82.50	22.69	77.08	24.91	*
Speech	94.17	20.77	92.50	24.05	-
Non-speech	70.83	39.37	61.66	43.54	*
Production time (sec)	15.88	3.68	15.70	4.46	-
Speech	15.52	4.18	15.33	4.58	-
Non-speech	16.24	5.07	16.06	5.79	-
Accuracy (%)	89.17	13.31	96.25	9.00	**
Speech	88.33	21.33	95.83	13.94	*
Non-speech	90.00	20.17	96.67	12.58	*
Reaction time (ms)	361	238	334	207	-
Speech	419	383	328	235	-
Non-speech	326	197	341	239	-

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

Errors resulting in existing words seem harder to identify than errors that result in non-existing words, both in the speech and non-speech condition. This result might also partly explain the difference in correction behavior (cf. delayed error correction).

5 Conclusions and discussion

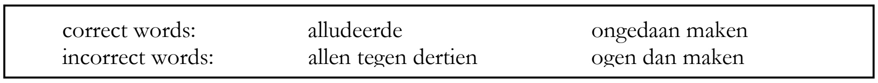
The experimental condition, in which the TPSF was either offered only on screen, or also read-aloud before the visual prompt, influences the writers' strategies during error analysis. In the speech condition writers delay error correction more often and start writing sooner than in the non-speech condition. When speech is present, writers can overtly compare the TPSF on screen with the speech. However, when speech is absent, only an internal, covert conflict is possible. The added speech sometimes confirms that the TPSF is the text intended and sometimes it does not. Without speech, this kind of explicit confirmation is not possible. The present results show that writers adjust to this uncertainty in the TPSF by correcting errors immediately and by needing more time to either continue text production or correct the error (hypothesis 1).

Error span has a quite consistent effect on strategy choice (hypothesis 2). The effect is especially powerful when interacting with the speech condition. A text that is not preceded by speech and that contains large errors leads to the highest accuracy of error correction. On the one hand large errors lead to longer preparation and production times, slower interference reaction times, indicating that they consume more working memory resources, and on the other hand they produce a higher rate of error analysis success than small errors. When confronted with a large error, writers seem to choose a wise speed accuracy tradeoff. And this successful strategy is associated with a pattern of correcting errors when they are detected at the point of utterance. One implication of this result in general is that the TPSF influences error analysis by encouraging frequent repairs of errors as they occur, even when they are large or more spread out in the text. The pausing behavior of writers provides extra evidence for the complexity of large errors. The pausing times were longer before large errors, indicating that it took writers more effort to solve them. However, when the error correction of the large errors was delayed the writers could easily return to the error, probably because they were so obvious to find. The correction was more difficult in this situation, perhaps because retrieving the correct information about the context is a highly demanding activity.

One benefit of the present research is that mode of writing can be separated from errors that are caused by a particular mode of writing, namely input mode. That is, writing with speech recognition generates a type of error that is distinct from those found when writing with keyboard & mouse, but the error types can be experimentally separated from mode of writing in the laboratory. In this experiment we have operationalized this on two levels: the input mode and the lexicality of words. The results for both operationalizations are to a large extent in line with each other (hypothesis 3 and 4). Both errors inherent to writing with speech recognition and with keyboard and mouse are equally distracting. But when speech recognition errors are involved the strategy is more likely to be associated with lower rates of error correction success and delay of correction. Keyboard based errors are better solved, especially when the TPSF is also read aloud.

We have seen that writers engage in different strategies to juggle the constraints that writing imposes on them. One of the biggest advantages of speech recognition over keyboard based word processing seems to be the fluency that writers can produce text with (respectively about 140 wpm versus 40 wpm). The main advantage of speech recognition over classical dictating is stated to be the immediate visual feedback on the screen. The TPSF is seen as an important input for further text generation. Concurrently the TPSF shows a delicate balance between cognitive processes. The re-reading of the last formulated sentences might have a positive effect on reducing working memory demands (Torrance & Gabraith, 2006). Alamargot, Dansac, Ros & Chuy (2005) showed recently that low-span writers read-back more frequently than high-span writers. However, Olive & Piolat (2002) found that removing visual feedback does not influence the quality of the text or the fluency with which it was written. The writers that were prevented from re-reading the short argumentative texts they were writing responded even faster to a secondary task. This provides evidence the hypothesis that re-reading the TPSF is very demanding. In our experiment the writers were in fact forced to read the TPSF, because they were asked to detect possible errors. So, the TPSF in general and the deficient text in particular appears to be a key element in the distribution of the working memory to free resources to do ‘a good job’ in any writing. Kellogg states that editing (in this study; repair) is a component of the central executive and reading of the central executive and the phonological loop[11]. This does not seem to be completely in line with the results of the current experiment since the visible TPSF on the screen seems to influence the strategies of writers. We would like to suggest integrating the visuo-spatial sketchpad in the description of the monitoring part of the writing model (cf. Figure 1). The visuo-spatial sketchpad is used to maintain and process visual or spatial information (Levy & Marek, 1999), for example visualizing ideas, organizational schemes and lay-out. Research showed that the visual working memory is also involved in composing concrete definitions (Sadowski, Kealy, Goetz, & Paivio, 1997). The spatial component helps in organizing ideas about the text hierarchically (Kellogg, 1988).

Since the relation between the monitoring process and the working memory is crucial in this study, we first relate our findings between monitoring and the phonological loop and consequently add the influence of the visuo-spatial sketchpad. As stated before, monitoring interacts with the central executive and the phonological loop. The results of our study confirm the role of the phonological loop. For instance, the preparation time – indicating the time the writer needed to read the context phrase that needed to be completed – is both influenced by the accuracy of the TPSF and by the speech recognition (e.g. the addition of speech shortens the preparation time to complete the sentence – preferred strategy). What is more, the editing strategy and, in most instances, also the reaction time seems to be influenced by these factors indicating a relation between the monitoring process and the phonological loop. However, the results also show a distinct relationship with the visuo-spatial sketchpad. One characteristic of speech recognition errors is that the different outputs are phonologically identical, but they might have a completely different visual representation. The phonological loop might be helpful in retrieving the correct word, but we see that large errors take more preparation time, also in the condition that is preceded with speech. According to us, the visuo-spatial sketchpad can be of importance in two aspects: the orthographic (re)presentation of the visual word stimulus (screen) and visual-tactile (keyboard). A large speech recognition error can be divergent from the intended output. Two Dutch examples are (in English respectively: allude (to) and



correct words:	alludeerde	ongedaan maken
incorrect words:	allen tegen dertien	ogen dan maken

[11] This topic in the discussion is partly based on personal communications with R. Kellogg of the University of St. Louis (June 2005 and February 2007). We would like to thank him for his supportive and critical remarks.

undo):

The mental picture of the correct word is rather deviant from the incorrect representation. In our opinion writers ‘see’ the incorrectness of the word, not only because they re-read the text, but also as an apparent visual misrepresentation. We assume this is a different kind of monitoring than purely reading for flaws in the text and generating new ideas. On the other hand, Kellogg states that ‘revising’[12] is a renewed activation of planning, translating, programming and execution. We believe the same holds for editing. However, the visuo-spatial component is not described yet from this point of view in the stages of error detection, diagnosing and correcting.
In this experiment the errors are already placed in the text, and the text appeared as if it were dictated. The visual-tactile line of approach is related to typing with keyboard (although one could compare it to a mispronunciation while dictating to a speech recognizer). According to the model a small typing error is handled solely by the central executive. Here too, the visual(-tactile) information may be significant. Blind typists and seeing typists might ‘feel’ the errors that they make. Typing and dictating requires a constant monitoring of the output on the screen. Again, writers verify the direct relation between the key pressed, the word dictated, with the output on the screen. This is not only demanding for the phonological loop, or the central executive, but also for the visuo-spatial sketchpad. The influence of the word-picture on error detection and the visual-tactile can be the object of further research.

6 Further research

In this study we introduced a new technique to experimentally isolate the impact of error type. The results show that our hypotheses are not always confirmed, and that the interaction of writers with the TPSF is a very complex process, especially when the intended context is not correctly represented. Of course, when interpreting these results we have to take into account certain limitations that are inherent to the experimental setting and the way in which we implemented the errors in our data set. In the last part of this article we would like to discuss some of these aspects and relate them to options for further research.

In testing the four hypotheses we consecutively compared the effect of different error types. However, the dataset of these separate error types is rather limited per category, mainly due to our choice to create a bias towards correct sentences and to explore the effect of different error types. Therefore, some of these results should be interpreted with caution and a replication with larger data sets per category might be a good way to validate the current results. At this stage, however, it was our major purpose to explore the effect of all the different error types. Previous research (Hacker et al., 1994; Larigauderie, Gaonac’h, & Lacroix, 1998; Leijten & Van Waes, 2005) did not provide enough information, nor on error types, nor on the effect of the writing mode. In a follow-up study we would like to use a more limited set of error types, allowing for a larger number of occurrences in the data set.

Since the data in this experiment are repeated measures – in line with the common practice in writing research – we chose to conduct GLM repeated measures[13] (ANOVA). However, from a statistical point of view, two objections can be made to our approach (De Maeyer & Rymenans, 2004; Quené & Van den Bergh, 2004). The first objection is that we did not take into account the nested data: the observations within each participant are correlated because they are made within the same participant. In other words, each sentence is nested in a participant. Since we aggregated the data to the participant level we did not take into account the hierarchical structure of the data (60 participants * 48 sentences should be 2880 observations: 1440 correct (*n* = 24), 1440 incorrect (*n* = 24)). In addition, we needed to code some data as missing, because a subject missed data on one observation. Since ANOVA does not allow missing data, more data needed to be discarded. Multilevel analysis has more statistical power. Therefore we will re-analyze the dataset conducting multilevel analyses.

Another aspect of this study is related to the experimental setting. In this study we simulated a writing context in which writers were confronted with the TPSF in different conditions. The writers needed to store (planning) information in their short term memory to correctly fulfill the writing (completion) task. Obviously, this task differs from a normal writing situation. For instance, the (re)-reading and writing process in the experiment is more explicitly focused on the sentence level and related to local planning behavior of writers; the context is not only created mentally, but presented visually as textual material. Since we isolated error types effects of error types

[12] Kellogg (1988) defines revising as a between draft activity: referring to the (final) phase of text composition.
[13] The same measurement is made several times on each subject or case. If between-subjects factors are specified, they divide the population into groups.

were described per type, while in complex writing tasks writers need to deal with combinations of error types. However, in spite of these differences, we observed similar patterns in the experimental setting as compared to our previous study in which we observed writers in their normal working environment (Leijten & Van Waes, 2005b). For instance, in both studies writers use different writing strategies to correct errors in the TPSF. Nevertheless, more research is needed – both ethnographically and experimentally – to validate the variety of cognitive processes during the interaction with the TPSF.

A third aspect of the study that needs further thought refers to the interpretation of certain subprocesses the writers are involved in during the interaction with the TPSF. This study confirms our hypothesis that errors in the TPSF influence the writing (and repair) strategies of writers. The data show that writers sometimes opt to correct the error first and sometimes continue with text production first. This difference is partially explained by the kind of error the writers are confronted with, and partially by personal preferences. In other words, a certain error type in the TPSF is likely to be handled in a specific way, but allows for certain variety. Based on the current study, however, it is difficult to relate this variety to certain personal preferences of the writers. Moreover, in those instances where the error in the TPSF is not corrected at the point of utterance, the collected data do not allow us to conclude whether the writers detected the error or not in the first interaction with the TPSF. This study gives insight in the global correction process of different types of errors, but to describe the error detection in more detail, it is necessary to replicate this study and add a more granulated observation method of the (re)reading behavior, for example eye-tracking. This technology would enable us to record every movement (saccade) and fixation of the writers' eye during the interaction with the TPSF. Data resulting from this kind of observation could provide a more detailed basis for analyzing the subprocess of error detection. For instance, when we observe a fixation of the eye on the error during the first interaction with the visual presentation of the TPSF, we will be able to differentiate more explicitly between 'postponing strategies' and those instances in which the error is simply overlooked (cf. Figure 1).

Another issue is closely related to this aspect of the observation methods involved in the study, and refers to the kinds of (re)reading involved in the interaction with the TPSF. During this interaction different types of reading might be involved. Rereading the TPSF is often aimed at building a context to produce new text. The focus of the reading process is on the pragmatic, semantic and/or syntactic aspects of the text to make the information in the short term memory more explicit. The interaction with the TPSF can either take place on a local level (word/phrase), or on a more global level (multiple sentences). However, during the interaction the focus of the reading might change. For instance, due to an error in the text the focus of the reading process might shift to a more evaluative type of reading paying more attention to the spelling or other correctness aspects (even at a local level). The writer's reading attention span may also be oriented towards a verification of the writer's intention. In the experimental setting we have tried to create a situation that allows for all types of reading described above. For instance, by presenting the participants more correct than incorrect phrases, we tried to create a default reading attitude which was not evaluative but one which was oriented towards the production of new text (i.c. completing the sentence with a causal pattern).

Nevertheless, in further research, we would like to be able to get more information about the different cognitive processes and orientation of the reading process. We hope that the registration of the eye-movements through eye-tracking can provide data that enable us to characterize the reading process on the local level in more detail. We think that a more refined analysis of the interaction with the TPSF is important for a better understanding of the dynamics underlying text production.

The final aspect refers to the correction profiles of the writers. As described in Figure 1, the participants in our case study (Leijten & Van Waes, 2005) differed in the way they preferred to repair errors during writing. This observation also holds for the participants in the experiment at hand. Some participants prefer to correct most errors before completing the text, while others do exactly the opposite, or change their strategy. The behavior of the last two groups in particular opens perspectives for further analysis. What is the rationale behind the strategy of delaying the correction of an error? Which cognitive aspects related to the working memory influence this behavior? Is the strategy dominated by an absolute preference, or is it related to the type of interaction with the TPSF and/or the type of error encountered during that interaction? A more refined analysis in which the data are studied from the perspective of the participants' profile and complemented with more detailed information of the cognitive (sub) process, might reveal complementary factors. This type of analysis could take into account different subprofiles of writing strategies that are used by writers who delay error repair.

Generally, five steps are characteristic for the interaction with the TPSF aimed at the formulation of new text: (a) reading of the TPSF – (b) detection of the error – (c) diagnosis – (d) editing – (e) completion of the text. Based on the walkthrough of these steps we can distinguish two diverse profiles: 'handle' versus 'postpone'.

Writers who do not delay the error correction follow this process linearly (ab-c d e), we would like to call this the ‘handle profile’; other writers prefer a non-linear process, that could be bundled in the terminology as a ‘postpone profile’. On the basis of our observations three different subprofiles can be distinguished in this last postpone group.

Gestalt profile [a e (a) bcd]: When following the ‘Gestalt profile’ writers do not really ‘read’ the TPSF, they only ‘perceive the Gestalt of the text’. They are mainly focused on grasping the main gist of the text in order to produce new text. Only after completing (part of) the text they detect and diagnose possible errors and decide to correct them.

Detection profile [ab e cd]: The ‘Detection profile’ is characterized by an initial reading process in which evaluation of the correctness of the TPSF takes place while reading. However, writers who follow this profile prefer not to diagnose and edit the error first. Probably because of constraints of the working memory (sometimes related to the complexity of the error), they decide to complete the text first and delay the diagnosis and correction.

Diagnosis profile [abc e d]: The ‘Diagnosis profile’ only differs from the ‘Detection profile’ because writers in this profile decide to delay the editing of the error after having diagnosed it.

Therefore, we would like to replicate this kind of study taking into account the handle and postpone profiles, if possible, in combination with a more refined analysis of the types of reading involved (cf. supra). We think that the variation of condition - with and without speech - might again add a more layered interpretation of the analyses. Further research should address this issue more explicitly.

Acknowledgements

We would like to thank Bart Van de Velde who did a great job in programming the experiment. We also would like to thank Michael Levy for adapting the Counting Span Test for Visual Basic.NET. Finally, we thank Tom Van Hout for proofreading this paper.

The project was funded as a BOF/NOI (New Research Initiatives) project by the University of Antwerp (2002-2004). The experiment software and logging tool are available from the internet at the following URLs respectively: www.ua.ac.be/marielle.leijten and www.inputlog.net.

References

Adams, J. W., Hitch, G., & Hutton, U. (1997). Working memory and children's mental addition. *Journal of Experimental Child Psychology, 67*, 21-38.

Alamargot, D., Dansac, C., Ros, C., & Chuy, M. (2005). Rédiger un texte procédural à partir de sources: Relations entre l'empan de production écrite et l'activité oculaire du scripteur. In D. Alamargot, P. Terrier & J. M. Cellier (Eds.), *Production, compréhension et usage des écrits techniques au travail* (pp. 51-68). Toulouse: Octarès.

Baddeley, A. D. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *Recent advances in learning and motivation* (Vol. 8, pp. 47-90). New York: Academic Press.

Blau, S. (1983). Invisible writing: Investigating cognitive processes in writing. *College, Composition and Communication, 34*, 297-312.

Bourdin, B., & Fayol, M. (1994). Is written language production more difficult than oral language production? A working memory approach. *International Journal of Psychology, 29*, 591-620.

Chenoweth, N. A., & Hayes, J. R. (2003). The Inner Voice in Writing. *Written communication, 20*(1), 99-118.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behaviour, 19*, 450-466.

De Maeyer, S., & Rymenans, R. (2004). *Onderzoek naar kenmerken van effectieve scholen: Kritische factoren in een onderzoek naar schooleffectiviteit in het technisch en beroepssecundair onderwijs in Vlaanderen. [Study of the characteristics of effective schools: Critical factors in school effectivity of secondary education in Flanders]* University of Antwerp, Antwerp.

Ericsson, K. A., & Kintsch, W. (1995). Long-Term Working Memory. *Psychological Review, 102*(2), 211-245.

Fisk, A. D., Derrick, W. L., & Schneider, W. (1986-87). A methodological assessment and evaluation of dual-task paradigms. *Current Psychology Research & Reviews, 5*, 315-327.

Gould, J. D. (1978). How experts dictate. *Journal of Experimental Psychology: Human Perception and Performance, 4*(4), 648-661.

Gould, J. D., & Alfaro, L. (1984). Revising documents with text editors, hand-writing recognition systems and speech-recognition systems. *Human Factors, 26*(4), 91-406.

Haas, C. (1989a). Does the medium make the difference? Two studies of writing with pen and paper and with computers. *Human-Computer Interaction, 10*, 149-169.

Haas, C. (1989b). How the writing medium shapes the writing process: Effects of word processing on planning. *Research in the Teaching of English, 23*, 181-207.

Hacker, D. J. (1994). Comprehension monitoring as a writing process. *Advances in Cognition and Educational Practice, 6*, 143-172.

Hacker, D. J. (1997). Comprehension monitoring of written discourse across early-to-middle adolescence. *Reading and Writing, 9*(3), 207-240.

Hacker, D. J., Plumb, C. S., Butterfield, E. C., Quathamer, D., & Heineken, E. (1994). Text revision: Detection and correction of errors. *Journal of Educational Psychology, 86*(1), 65-78.

Honeycutt, L. (2003). Researching the use of voice recognition writing software. *Computers and Composition, 20*, 77-95.

Hoskyn, M., & Swanson, H. (2003). The relationship between working memory and writing in younger and older adults. *Reading and Writing, 16*, 759-784.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in Working Memory. *Psychological Review, 99*(1), 122-149.

Kellogg, R. T. (1988). Attentional overload and writing performance: Effects of rough draft and outline strategies. *Journal of Experimental Psychology: Learning, memory and cognition, 14*(2), 355-365.

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. E. Ransdell (Eds.), *The Science of Writing: Theories, methods, individual differences and applications* (pp. 57-71). Hillsdale, NJ: Lawrence Erlbaum.

Kellogg, R. T. (1999). Components of Working Memory in Text Production. In M. Torrance & G. Jeffery (Eds.), *The cognitive demands of writing: processing capacity and Working Memory effects in text production* (Vol. 3, pp. 43-61). Amsterdam: Amsterdam University Press.

Kellogg, R. T. (2001). Competition for working memory among writing processes. *American Journal of Psychology, 114*, 175-191.

Kellogg, R. T. (2004). Working memory components in written sentence generation. *American Journal of Psychology, 117*, 341-361.

Larigauderie, P., Gaonac'h, D., & Lacroix, N. (1998). Working memory and error detection in texts: What are the roles of the central executive and the phonological loop? *Applied Cognitive Psychology, 12*, 505-527.

Leijten, M., & Van Waes, L. (2005). Writing with speech recognition: The adaptation process of professional writers with and without dictating experience. *Interacting with Computers, 17*(6), 736-772.

Leijten, M., & Van Waes, L. (2006). Repair strategies in writing with speech recognition: The effect of experience with classical dictating. In L. Van Waes, M. Leijten & C. Neuwirth (Eds.), *Writing and Digital Media* (Vol. 17, pp. 31-46). Oxford: Elsevier.

Levy, C. M., & Marek, P. (1999). Testing components of Kellogg's multicomponent models of Working Memory in writing: The role of the phonological loop. In M. Torrance & G. Jeffery (Eds.), *The cognitive demands of writing. Processing capacity and Working Memory effects in text production.* (Vol. 3, pp. 25-41). Amsterdam: Amsterdam University Press.

Levy, C. M., & Ransdell, S. E. (2001). Writing with concurrent memory loads. In T. Olive & C. M. Levy (Eds.), *Contemporary Tools and Techniques for Studying Writing* (pp. 9-29). Dordrecht: Kluwer Academic Publishers.

Lindgren, E., & Sullivan, K. P. H. (2006a). Analysing on-line revision. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer Keystroke Logging: Methods and Applications* (Vol. 18, pp. 157-188). Oxford: Elsevier.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8*(3), 299-325.

Olive, T. (2004). Memory in writing: Empirical evidences from the dual-task technique working. *European Psychologist, 9*(1), 32-42

Olive, T., & Kellogg, R. T. (2002). Concurrent activation of high- and low-level production processes in written composition. *Memory and Cognition, 30*, 594-600.

Olive, T., & Piolat, A. (2002). Suppressing visual feedback in written composition : Effects on processing demands and coordination of the writing processes. *International Journal of Psychology, 37*(4), 209-218.

Piolat, A., Roussey, J. Y., Olive, T., & Amada, M. (2004). Processing time and cognitive effort in revision: effects of error type and of working memory capacity. In L. Allal, L. Chanquoy, P. Largy & Y. Rouiller (Eds.), *Revision: Cognitive and Instructional Processes* (pp. 21–38). Dordrecht: Kluwer Academic Publishers.

Quené, H., & Van den Bergh, H. (2004). On Multi-Level Modeling of data from repeated measures designs: A tutorial. *Speech Communication, 43*(1-2), 103-121.

Quinlan, T. (2004). Speech recognition technology and students with writing difficulties: Improving fluency. *Journal of Educational Psychology, 96*, 337-346.

Quinlan, T. (2006). Young Writers and Digital Scribes. In L. Van Waes, M. Leijten & C. Neuwirth (Eds.), *Writing and Digital Media* (Vol. 17, pp. 21-29). Oxford: Elsevier.

Rabbitt, P. (1978). Detection of errors by skilled typists. *Ergonomics, 21*, 945-958.

Rabbitt, P., Cummings, P., & Vyas, S. (1978). Some errors of perceptual analysis in visual search can be detected and corrected. *Quarterly Journal of Experimental Psychology, 30*, 417-427.

Ransdell, S. E., & Hecht, S. A. (2003). Time and resource limits on working memory: Cross-age consistency in counting span performance. *Journal of Experimental Child Psychology, 86*, 303-313.

Ransdell, S. E., & Levy, C. M. (1996). Working memory constraints on writing quality and fluency. In C. M. Levy & S. E. Ransdell (Eds.), *The science of writing: Theories, Methods, Individual Differences, and Applications* (pp. 93-105). Mahwah, NJ: Lawrence Erlbaum Associates.

Ransdell, S. E., & Levy, C. M. (1999). Writing reading and speaking memory spans and the importance of resource flexibility. In M. Torrance & G. Jeffery (Eds.), *The cognitive demands of writing: processing capacity and working memory effects in text production*. Amsterdam: Amsterdam University Press.

Ransdell, S. E., Levy, C. M., & Kellogg, R. T. (2002). The structure of writing processes as revealed by secondary task deman. *L-1-Educational Studies in Language and Literature 2*(2), 141-163.

Sadowski, M., Kealy, W. A., Goetz, E. T., & Paivio, A. (1997). Concreteness and imagery effects in the written composition of definitions. *Journal of Experimental Psychology., 89*, 518-526.

Schilperoord, J. (1996). *It's about time: Temporal aspects of cognitive processes in text production*. Amsterdam/Atlanta: Rodopi.

Severinson Eklundh, K. S. (1994). Linear and Non-linear strategies in computer-based writing. *Computers and Composition, 11*, 203-216.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General, 125*, 4-27.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders's method. *Acta Psychologica, 30*, 276-235.

Torrance, M., & Galbraith, D. (2006). The processing demands of writing. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 468). New York: Guilford Publications.

Towse, J., Hitch, G., & Hutton, U. (1998). A reevaluation of working memory capacity in children. *Journal of Memory and Language, 39*, 195-217.

Van den Bergh, H., & Rijlaarsdam, G. (1996). The dynamics of composing: Modelling writing process data. In C. M. Levy & S. E. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 207-232). Mahwah, NJ: Lawrence Erlbaum Associates.

Van Waes, L., & Schellens, P. J. (2003). Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers. *Journal of Pragmatics, 35*(6), 829-853.

Appendix 1

Example of four kinds of errors

<table>
<tr><th colspan="2">Large difference, occuring only in speech recognition writing mode</th></tr>
<tr><td>context</td><td>Omdat niemand anders het durft, wil ik zeggen dat het vak interessant is.
Because no one else dares, I would like to say that the course is interesting.</td></tr>
<tr><td>1st segment</td><td>Dat wil zeggen dat het vak eendjes Zand is, want
That will say that the course duckling Sand is, because (literal)
That means that the course is in the rest thing, because (plausible)</td></tr>
<tr><th colspan="2">Small difference, occuring in speech recognition</th></tr>
<tr><td>context</td><td>Omdat hij mijn resultaten wilde bekijken, haalde hij het verslag uit mijn kast.
Because he wanted to take a look at my results, he took the report out of my closet.</td></tr>
<tr><td>1st segment</td><td>Hij haalde het verslag uit mij kast, want
He took the report out of my closet, because</td></tr>
<tr><th colspan="2">Small difference, occcuring in speech recognition and keyboard & mouse writing mode
(existing words)</th></tr>
<tr><td>context</td><td>Omdat je in het buitenland zat, ben je niet uitgenodigd op dat feestje.
Because you were abroad you were not invited to that party.</td></tr>
<tr><td>1st segment</td><td>Je vent niet uitgenodigd op dat feestje, want
You bloke not invited to that party, because <i>(literal)</i>
You art not invited to that party, because <i>(plausible)</i></td></tr>
<tr><th colspan="2">Small difference, only occuring in keyboard & mouse writing mode (non-existing words)</th></tr>
<tr><td>context</td><td>Omdat Jan tegen een boom was gereden, had de fiets schade aan het voorwiel.
Because John had crashed into a tree, the bike’s front wheel was damaged.</td></tr>
<tr><td>1st segment</td><td>De fiets had schade aan het voorwiem*, because
De bike was damaged at the front wheek, because

<i>* The keys ‘l’ and ‘m’ are adjacent keys on a keyboard with azerty settings</i></td></tr>
</table>

Appendix 2

Overview of instructions

General procedure

Course of the experiment
The session will last about one hour and consists of 5 parts.

1. Memory test
2. Reaction test
3. Writing test | part I
4. Writing test | part II
5. Questionnaire

You will receive a short instruction before each test. Then you get the opportunity to practice. There is a short break after each test.

Good luck.

Counting span test

Memory test
The memory test shows a card with blue squares and red triangles.
Count the blue squares.
Enter the number of blue squares as soon as possible.
Hereafter appears a next page with squares and triangles.
Repeat the procedure.
Remember the number of blue squares of the previous pages.
After 2, 3 or 4 times you enter the numbers again in the boxes that appear.

The edges of the cards show you how many numbers you need to remember. In the example below it is three.

The memory test consists of three sessions: one practice session and two basic sessions.

To start with the practice session click ‘ok’.

Introduction Baseline Reaction Test

Reaction test
In this test we will measure your ability to react.
Wait until you hear a beep.
React as quickly as possible to the beep by pressing the button next to the computer.
Put you hands back to your keyboard afterwards.

Don’t keep your hands on the button.

Procedure 1: non-speech condition

Writing test | part I

A sentence will appear on the screen.

Read this sentence carefully and press the ‘ok’ button when you are finished.

A new sentence appears on the screen.

Complete this sentence as quickly as possible with the information from the first sentence. Some sentences that you need to complete might contain an error. Correct the error as soon as possible. You are free to choose whether you prefer to complete the sentence first or whether you correct the error first.

Remark:

The sentence needs to be perfect before you continue with the next sentence.

Attention please!

Now and then you will hear a beep. Push the button next to the computer as soon as possible. The first sentences that will appear are test sentences. This way you can practice.

Procedure 2: speech condition

Writing test | part II

A sentence will appear on the screen.

Read this sentence carefully and press the ‘ok’ button when you are finished.

You will hear a part of the next sentence via the headset.

A new sentence appears on the screen.

Complete this sentence as quickly as possible with the information from the first sentence. Some sentences that you need to complete might contain an error. Correct the error as soon as possible. You are free to choose whether you prefer to complete the sentence first or whether you correct the error first.

Remark:

The sentence needs to be perfect before you continue with the next sentence.

Attention please!

Now and then you will hear a beep. Push the button next to the computer as soon as possible. The first sentences that will appear are test sentences. This way you can practice.